Eine Hauptkomponenten-, Faktoren- und Clusteranalyse der Wirtschaftsnachrichten von Mashable.com

A Principal Component Analysis, Factor Analysis and Cluster Analysis of the Business News of Mashable.com

Bachelorarbeit zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.) in Betriebswirtschaftslehre
an der Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

vorgelegt von

Oliver Brose

Immatrikulationsnummer: 552147

Erstgutachter: Prof. Dr. Härdle

Zweitgutachterin: Prof. Dr. Müller

Betreuer: Dr. Klinke

Inhaltsverzeichnis

1 Einleitung

Im Juli 2005 gründete der damals 19-jährige Pete Cashmore das Online-Nachrichtenportal Mashable (www.mashable.com) (Beier und Wolfman 2012), auf dem Artikel über Themen wie Wirtschaft, Lifestyle, Social Media und Unterhaltung veröffentlicht werden. Die Leserzahlen wuchsen schnell - von 2 Millionen monatlichen Lesern im Januar 2007 auf 35 Millionen monatliche Leser im Oktober 2014 (Beier und Wolfman 2012; Peterson-Withorn 2014). Ein Grund für den Erfolg könnte die umfangreiche Lese- und Designoptimierung der Texte sein. So wird berichtet (Raymond 2013), dass bei Mashable mit erheblichem Aufwand verschiedene Überschriften, Textlängen und Designstrategien gegeneinander getestet werden, um Leser länger auf der Website zu halten - und sie anzuregen, die Artikel öfter in den sozialen Netzwerken (z.B. Facebook) zu verbreiten. Dadurch können kostenlos neue Leser gewonnen werden.

Doch welche Faktoren führen dazu, dass manche Artikel von Mashable in sozialen Netzen von den Lesern häufig verbreitet werden, während andere Artikel von Mashable weniger häufig verbreitet werden? Um diese Frage zu beantworten, erhoben und untersuchten Fernandes, Vinagre und Cortez (2015a) einen Datensatz aus 39644 Artikeln, welche von Mashable veröffentlicht wurden. Der Datensatz wurde nach der Analyse der Öffentlichkeit zur Verfügung gestellt (Fernandes, Vinagre und Cortez 2015b).

Es gibt eine Reihe von Arbeiten (z.B. Dumont, Macdonald und Rincón; Lei und Hu), in denen diese Daten weiter untersucht wurden. In diesen Arbeiten untersuchte man, wie man die Verbreitung der Artikel in sozialen Netzwerken vorhersagen kann. Gruppierungen von ähnlichen Artikeln oder Beziehungen zwischen den erhobenen Variablen wurden jedoch noch nicht tiefergehend untersucht. Diese offenen Themen sollen in dieser Arbeit untersucht werden. Konkret sollen dabei diese Fragen bearbeitet werden:

1. Kann man Variablen des Datensatzes effizient zusammenfassen (Variablenreduktion bzw. Dimensionsreduktion) und stehen hinter manchen Variablen nichtbeobachtbare (latente) oder nicht erfasste Faktoren? Wie ist die Beziehung zwischen den erfassten Variablen?
2. Gibt es Gruppierungen von ähnlichen Artikeln (z.B. Gruppen mit ähnlichem Wortschatz und Sprachstil)? Wenn ja – was zeichnet diese aus?

Um diese Fragen zu beantworten, wird eine Korrelationsanalyse, eine Hauptkomponentenanalyse, eine explorative Faktorenanalyse und eine Clusteranalyse durchgeführt. Um diese Untersuchung (und ihre Ergebnisse) potentiell zu fokussieren, werden dabei zwei Einschränkungen vorgenommen:

1. Es werden ausschließlich die Wirtschaftsnachrichten untersucht. Der Grund dafür ist, dass Mashable Artikel zu vielen unterschiedlichen Themengebieten (z.B. Lifestyle, Unterhaltung und Wirtschaft) veröffentlicht, die sich in Variablen (wie z.B. Wortschatz oder Sprachstil) potentiell stark unterscheiden können - was dazu führen würde, dass die Ergebnisse der Hauptkomponenten-, explorativen Faktoren- und Clusteranalyse themenabhängig sein können. Eine themenabhängige Analyse scheint damit angebracht. Die Wirtschaftsnachrichten wurden ausgewählt, da diese die Sicht der Leser auf Politik, Wirtschaft und Unternehmen beeinflussen können (Mutz und Soss 1997; Kiousis, Popescu und Mitrook 2007; Iyengar und Kinder 2010) und damit als gesellschaftspolitisch besonders relevant betrachtet werden können.
2. Es werden nur Variablen untersucht, die eine erkennbare Auswirkung auf Inhalt und den Stil eines Artikels haben (z.B. Wortschatz, Subjektivität, Anzahl der Bilder, …). Einige technische Kennzahlen - zum Beispiel zur Suchmaschinenoptimierung - werden in dieser Arbeit nicht beachtet, da *die Texte selbst* (anhand ihrer Kennzahlen) untersucht werden sollen. Viele technische Kennzahlen sind für den Leser nicht sichtbar und haben für das Leseerlebnis folglich eine untergeordnete Rolle.

Die Arbeit ist wie folgt aufgebaut: Zuerst werden die Variablen und der Datensatz, die Datenerhebung, die Datenverarbeitung und die Variablenauswahl beschrieben. Danach werden die Methoden – die Korrelationsanalyse, die Hauptkomponentenanalyse, die Faktorenanalyse und die Clusteranalyse – beschrieben. Es wird erklärt und begründet, wie diese Methoden im praktischen Teil der Arbeit angewendet werden. Im darauffolgenden Teil der Arbeit werden die Verfahren angewendet und ihre Ergebnisse interpretiert und diskutiert. Im letzten Teil der Arbeit werden die Daten, die Analyse und ihre Ergebnisse zusammengefasst und kritisch betrachtet.

2 Datensatz und Variablen

2.1 Datenerhebung, Grundgesamtheit und Stichprobentyp

Im Erhebungszeitraum vom 07.01.2013 bis 07.01.2015 (Fernandes et al. 2015a, S. 537 ff.) wurden Daten von insgesamt 39644 Artikeln der Webseite www.mashable.com mit Hilfe einer Software automatisiert extrahiert und verarbeitet. Es wurden jedoch nicht alle Artikel, welche im Erhebungszeitraum veröffentlicht wurden, extrahiert. So konnte zum Beispiel die Datenstruktur von einigen Artikeln nicht von der Software automatisiert verarbeitet werden (Fernandes et al. 2015a, S. 537 ff.). Darüber hinaus wurden einige Artikel von der Untersuchung ausgeschlossen, welche nur wenige Tage vor dem 07.01.2015 veröffentlicht wurden. Für

diese Artikel wäre ihrer Aussage nach nicht genug Zeit gewesen, von Lesern in sozialen Netzwerken verbreitet zu werden. Fernandes et al. (2015a) geben an, dass nur ein kleiner Teil der im Untersuchungszeitraum veröffentlichten Artikel von diesen Einschränkungen betroffen ist - sie nennen jedoch keine konkrete Zahl.

Die Grundgesamtheit der Untersuchtung lautet wie folgt: „Alle Artikel, welche zwischen dem 07.01.2013 und 07.01.2015 auf mashable.com veröffentlicht wurden". Aufgrund der Datenerhebung handelt es sich bei der Stichprobe um eine *bewusste Auswahl*. Bei dieser wird nur ein Teil der Beobachtungen einer Grundgesamtheit erfasst, welcher anhand willkürlich festgelegter Kriterien ausgewählt wird (Schnell, Hill und Esser 2011, S. 259). Die willkürlich festgelegten Kriterien sind in diesem Fall zwei Dinge: Die Extraktion hat nur für Artikel mit einer ausgesuchten Dateistruktur stattgefunden, und es wurden einige Artikel aussortiert, die kurz vor dem Ende des Erhebungszeitraums in den sozialen Netzwerken nicht ausreichend verbreitet werden konnten. Eine Schlussfolgerung der Stichprobe auf die Gesamtheit der auf Mashable veröffentlichten Artikel (die Grundgesamtheit) ist durch das Stichprobenverfahren nicht möglich (Schnell et al. 2011, S. 260).

2.2 Daten, Variablenbeschreibung und Variablenauswahl

Aus den Artikeln haben die Autoren 60 analysierbare Variablen gewonnen, welche durch eigene Berechnungen (Fernandes et al. 2015a) und die Software „Pattern" (De Smedt und Daelemans 2012) erzeugt wurden. Die Variablen sind zur besseren Übersicht in dieser Arbeit in sechs Klassen eingeteilt: „Sentimentanalyse", „Deskriptive Textbeschreibung", „Thema", „Erhebung", „technische Kennzahlen" und „Veröffentlichung".

Variablen der Kategorien „Erhebung", „technische Kennzahlen" und „Veröffentlichung" werden in der Analyse dieser Arbeit, wie in der Einleitung bereits beschrieben, nicht betrachtet (aber im Anhang beschrieben). Variablen der anderen drei Klassen werden in der Analyse dieser Arbeit verwendet und deshalb an dieser Stelle beschrieben:

Variablenklasse „Sentimentanalyse". Das Ziel der Sentimentalanalyse ist es, Meinungen und Stimmungen aus Texten mit Hilfe von Computern zu extrahieren (Liu 2015, S. XIV). Um dieses Ziel zu erreichen, kann man unter anderem Kennzahlen der Polarität und Subjektivität berechnen (Liu 2010, S. 1 ff.). In der Literatur gibt es für die Begriffe Polarität und Subjektivität keine einheitlichen Definitionen, aber eine allgemeine Auffassung (Montoyo, Martínez-Barco und Balahur 2012; Przepiórkowski und Ogrodniczuk 2014; Pang und Lee 2008), welche durch diese Definitionen wiedergegeben wird:

1. Polarität ist in der Sentimentalanalyse eine Unterteilung in negative, neutrale und positive Stimmung der Wörter (Liu 2015, S. 11).
2. Subjektivität kann man als (polare) innere Zustände, welche nicht offen sind für Fakten - bzw. nicht auf diesen basieren - definieren (Przepiórkowski und Ogrodniczuk 2014, S. 313; Liu 2010, S. 1 ff.).

Variablenkürzel	Variablenbeschreibung	Wertebereich
global_subjectivity	Subjektivitätswert des gesamten Artikels	[0,1]
global_sentiment_polarity	Polaritätswert des gesamten Artikels	[-1,1]
global_rate_positive_words	Anteil der positiven Wörter von allen Wörtern im Artikel	[0,1]
global_rate_negative_words	Anteil der negativen Wörter von allen Wörtern im Artikel	[0,1]
rate_positive_words	Anteil an positiven Wörter von allen nichtneutralen Wörtern	[0,1]
rate_negative_words	Anteil an negativen Wörtern von allen nichtneutralen Wörtern	[0,1]
avg_positive_polarity	Durchschnittliche Polarität der positiven Wörter	[0,1]
min_positive_polarity	Minimaler Polaritätswert der positiven Wörter	[0,1]
max_positive_polarity	Maximaler Polaritätswert der positiven Wörter	[0,1]
avg_negative_polarity	Durchschnittliche Polarität der negativen Wörter	[-1,0]
min_negative_polarity	Minimaler Polaritätswert der negativen Wörter	[-1,0]
max_negative_polarity	Maximaler Polaritätswert der negativen Wörter	[-1,0]
title_subjectivity	Subjektivitätswert der Überschrift	[0,1]
title_sentiment_polarity	Polaritätswert der Überschrift	[-1,1]
abs_title_subjectivity	Absolute Differenz des Subjektivitätswerts der Überschrift zu 0,5	[0;0.5]
abs_title_sentiment_polarity	Absolute Differenz des Polaritätswert der Überschrift zu 0,5	[0;1,5]

Tabelle 1: Variablenbeschreibung und Wertebereich für die metrisch skalierten Variablen der Klasse „Sentimentalanalyse“. Quelle: Eigene Darstellung.

Deskriptive Textbeschreibung. Variablen der Kategorie „Deskriptive Textbeschreibung“ beschreiben gewisse Aspekte des Stils und Aufbaus der Artikel, gehören jedoch nicht zu den Kennzahlen einer Sentimentalanalyse. Beispiele für solche Kennzahlen können Wortschatz und Anzahl von Bildern in Artikeln sein. Die genauere Erläuterung der Variablen dieser Variablenklasse erfolgt in Tabelle 2. Zwei Begriffserklärungen sind zum Verständnis einiger Variablen dieser Variablenklasse wichtig:

Token. Ein Token ist eine Kette von Zeichen (z.B. Buchstaben und Zahlen), die durch ein Trennkriterium – z.B. ein Leerzeichen, Bindestrich oder Apostroph - unterbrochen wird (Carstensen, Ebert, Ebert, Jekat, Langer und Klabunde 2009, S. 264 ff.). Der Satz „Das ist ein Token.“ enthält zum Beispiel vier Tokens, der Name „Lena Meyer-Landruth“ enthält zwei Tokens, wenn das Trennkriterium ein Leerzeichen ist - und drei Tokens, wenn sowohl das Leerzeichen, als auch ein Bindestrich als Trennkriterium festgelegt wird. Die Autoren geben in der Publikation nicht an, welches Trennkriterium zur Datenverarbeitung verwendet wurde.

Stoppwort. Ein Stoppwort ist ein Wort, welches für den Inhalt eines Textes keine wichtige Bedeutung hat, aber in Texten häufig vorkommt (Goker und Davies 2009, S. 237). Beispiele für Stoppwörter sind „der“, „die“ und „das“.

Variablenkürzel	Variablenbeschreibung	Wertebereich
n_tokens_title	Anzahl der Tokens in der Überschrift eines Artikels	[0,a]
n_tokens_content	Anzahl der Tokens eines Artikels	[0,a]
n_unique_tokens	Anzahl einzigartiger Tokens	[0,a]
n_non_stop_words	Anzahl an Wörtern im Artikel, die keine Stoppwörter sind	[0,a]
n_non_stop_words	Anzahl an Wörtern im Artikel, die keine Stoppwörter sind	[0,a]
n_non_stop_unique_tokens	Anzahl an einzigartigen Tokens, die keine Stoppwörter sind	[0,a]
num_hrefs	Anzahl der Hyperlinks im Artikel	[0,a]
num_self_hrefs	Anzahl der auf Mashable verweisenden Links	[0,a]
num_imgs	Anzahl der Bilder	[0,a]
num_videos	Anzahl der Videos	[0,a]
average_token_length	Durchschnittliche Zeichenanzahl der Token	[0,b]

Tabelle 2: Variablenbeschreibung und Wertebereich für die Variablen der Klasse Textbeschreibung. Die Variable *a* ist Element der positiven natürlichen Zahlen, die Variable *b* ist Element der positiven reellen Zahlen. Quelle: Eigene Darstellung

Thema. Mit Hilfe einer latenten Dirichlet Allokation (LDA) nach Blei, Ng und Jordan (2003) wurden aus den Artikeln fünf Themen herausgearbeitet und die Nähe der Artikel zu diesen Themen kalkuliert. Die LDA findet in Texten eine vom Benutzer vorgegebene Zahl unterschiedlicher Themengebiete und berechnet die Nähe des Textes zu diesen Themengebieten (0 steht für maximal entfernt und 1 für maximal nah).
Fernandes et al. (2015a) haben nicht begründet, weshalb sie fünf Themen herausgearbeitet haben (warum zum Beispiel nicht sechs?). Sie haben auch nicht beschrieben, was die fünf Themen inhaltlich ausmacht. Man kann daher lediglich sagen, wie weit ein Artikel von den einzelnen Themen entfernt ist - nicht aber, was die einzelnen Themen konkret bedeuten.

Variablenkürzel	Variablenbeschreibung	Wertebereich
LDA_00	Gibt an, wie stark der Artikel Thema 0 zugeordnet ist	[0,1]
LDA_01	Gibt an, wie stark der Artikel Thema 1 zugeordnet ist	[0,1]
LDA_02	Gibt an, wie stark der Artikel Thema 2 zugeordnet ist	[0,1]
LDA_03	Gibt an, wie stark der Artikel Thema 3 zugeordnet ist	[0,1]
LDA_04	Gibt an wie stark der Artikel Thema 4 zugeordnet ist	[0,1]

Tabelle 3: Variablenbeschreibung und Wertebereich für die metrisch stetig skalierten Variablen der Klasse „Thema“. Quelle: Eigene Darstellung.

3. Methoden

3.1 Korrelationsanalyse

Der Korrelationskoeffizient nach Bravais-Pearson (folgend Korrelation genannt) ist eine dimensionslose Kennzahl für die Ermittlung des linearen Zusammenhangs zwischen zwei metrisch skalierten Variablen, welche Werte zwischen minus eins (perfekte negative Korrelation) und eins (perfekte positive Korrelation) annehmen kann (Cleff 2011, S. 107 ff.). Negative Korrelation bedeutet, dass die Werte der einen Variable steigen, wenn die der anderen fallen und positive Korrelation bedeutet, dass die Werte der einen Variable steigen, wenn die der anderen steigen (Cleff 2011, S. 109 ff.). Liegt die Korrelation im Bereich nahe null, existiert kein linearer Zusammenhang zwischen den Variablen. Die Formel lautet (Klinke 2014):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Die Variable r_{xy} steht dabei für die Korrelation r zwischen den Beobachtungen x und y. Wenn ein Datensatz Variablen besitzt, kann man in einer $p \times p$ Korrelationsmatrix kompakt alle Korrelationen zwischen jedem Paar der p Variablen darstellen. Die Korrelation r_{ij} ist dann, als Element der Korrelationsmatrix, die Korrelation zwischen der *i-ten* und *j-ten* Variable ($i,j=1, \ldots,p$).

3.2 Explorative Faktoranalyse

3.2.1 Ziele und Methodische Einordnung

Mit Hilfe der Faktorenanalyse formuliert man ein statistisches Modell, um die p Variablen eines Datensatzes durch q Faktoren abzubilden, wobei $q<p$ gelten soll (Härdle und Simar 2003, S. 275). Durch dieses statistische Modell kann man die Zahl der Variablen reduzieren; man kann nichtbeobachtbare (latente) Faktoren berechnen, und man kann korrelierte Variablen im Datensatz (Multikollinearität) entdecken und auflösen (Brown, Onsman und Williams, 2012).

Die Faktorenanalyse ist ein Spezialfall eines latenten Variablenmodells - es wird ein Zusammenhang zwischen beobachtbaren und nichtbeobachtbaren Variablen hergestellt - bei dem beobachtbare und nichtbeobachtbare Variablen metrisch skaliert sind (Galbraith, Moustaki, Bartholomew und Steele, 2002, S. 143 ff.). Brown et al. (2012) unterteilen die Faktorenanalyse weiterhin in die explorative und konfirmatorische Faktorenanalyse.

Bei der explorativen Faktorenanalyse trifft man keine Annahmen über Zusammenhänge zwischen Variablen, Variablen und Faktoren oder der Zahl der Faktoren – und wenn man solche Annahmen trifft, fließen diese nicht in die Ermittlung des statistischen Modells ein. Bei der

konfirmatorischen Faktorenanalyse trifft man dagegen Annahmen über Zusammenhänge zwischen Variablen, Variablen und Faktoren oder der Zahl der Faktoren und lässt diese in das statistische Modell einfließen.
In dieser Arbeit wird die explorative Faktorenanalyse verwendet, da vor der Untersuchung keinerlei Annahmen über Zusammenhänge zwischen Variablen, Variablen und Faktoren oder der Zahl der Faktoren gemacht werden.

3.2.2 Das orthogonale Faktorenmodell: Funktionsprinzip und Annahmen

Wenn man eine Beobachtung x mit p Variablen als Zufallsvektor $x=(x_1, \ldots, x_j, \ldots, x_p)^T$ betrachtet, von welchem alle Elemente den Erwartungswert Null haben, kann man ein Element x_j durch folgende lineare Gleichung darstellen (Härdle und Simar 2003, S. 276):

$$x_j = \lambda_{j1} f_1 + \ldots + \lambda_{jl} f_l + \ldots + \lambda_{jq} f_q + u_j \text{ mit } l=1, \ldots, q$$

Die Variable q gibt dabei die Zahl der Faktoren f_l an, welche weitaus kleiner sein soll als p. Die Variable u_j steht für den Teil von x_j, welcher nicht durch λ_{jl} und f_l erklärt werden kann. In Matrixschreibweise kann man dieses Modell wie folgt definieren:

$$\underset{(p\times 1)}{X} = \underset{(p\times q)}{\Lambda} \underset{(q\times 1)}{F} + \underset{(p\times 1)}{U}$$

Λ ist eine nichtzufällige Matrix und enthält k Faktorladungen λ_{jl}, welche angeben, wie groß der lineare Einfluss des Faktors f_l auf das Element xj ist. F ist ein nichtbeobachtbarer Zufallsvektor und enthält die q Faktorwerte f_l, die den Wert angeben, den eine Beobachtung auf diesem Faktor einnimmt. U ist ein nichtbeobachtbarer Zufallsvektor und erfasst den Teil von X, der nicht durch ΛF erklärt werden kann. An die Gleichungen müssen mehrere Annahmen gelten, damit es die Bedingungen des orthogonalen Faktorenmodells erfüllt (Härdle und Simar 2003 S. 277):

1. Die Fehlerterme U haben den Erwartungswert Null: $E[U_i]=0 \text{ mit } i=1, \ldots, p$.
2. Die Fehlerterme U sind untereinander unkorreliert: $E[U_i, U_j]=0 \text{ für } i \neq j \text{ mit } i,j = 1, \ldots, p$.
3. Die Varianzen ψ_{ii} zwischen den Fehlertermen U können unterschiedlich groß (heteroskedastisch) sein: $E[U_i, U_i]=\psi_{ii} \text{ mit } i= 1, \ldots, p$.
4. Faktoren und Fehlerterme sind unkorreliert: $Cov[F,U]=0$.
5. Die Faktoren F haben den Erwartungswert Null: $E[F_i]=0 \text{ mit } i=1, \ldots, q$.
6. Die Varianzen von F sind 1: $Var[F]=I_q$.
7. Die Faktoren sind unkorreliert: $Cov[F,F]=0$.

Die Varianz von x_j lässt sich im orthogonalen Faktorenmodell wie folgt zerlegen:

$$Var(x_j) = \lambda^2_{11} + \ldots + \lambda^2_{ql} + \psi_{jj} = h^2_j + \psi_{jj} \text{ mit } l=1, \ldots, q$$

Die Variable h_j^2 heißt Kommunalität und gibt an, welcher Wert der Varianz einer Variablen des Datensatzes durch das Faktorenmodell erklärt wird. Je höher h_j^2 und je kleiner ψ_{jj}, umso besser erklärt das Faktorenmodell die Variablen.

3.2.3 Ablauf der explorativen Faktorenanalyse in der Praxis

Brown et al. (2012) schlagen vor, eine explorative Faktorenanalyse in fünf Schritten durchzuführen:

1. Anforderungen an die Daten überprüfen
2. Extraktionsmethode auswählen
3. Faktoren extrahieren und Faktoren auswählen
4. Rotationsmethode auswählen
5. Faktoren interpretieren und benennen

Schritt 1: Anforderungen an die Daten überprüfen. Damit aus einer Faktorenanalyse verwertbare Ergebnisse erzeugt werden können, müssen mehrere Anforderungen an den Datensatz erfüllt sein. So müssen die Beobachtungen voneinander unabhängig sein, beobachtbare und nichtbeobachtbare Variablen müssen metrisch skaliert sein (Klinke 2015) und die Beziehungen zwischen den Variablen müssen linear sein (Yalcin und Amemiya 2001). Es gibt auch Anforderungen an den Mindeststichprobenumfang, wobei die Meinungen dazu in der Literatur weit auseinandergehen (Brown et al. 2012; Smith und Zygmont 2014; Costello und Osborn 2005): So werden manchmal Mindeststichprobenumfänge zwischen $n=100$ und $n=300$ oder Qualitätsabstufungen von $n=100$ (schlecht) bis $n=1000$ (exzellent) vorgeschlagen, und manchmal werden Variablen:Beobachtungen-Verhältnisse von 3:1 bis 20:1 vorgeschlagen. Hong, MacCallum, Widaman und Hong (1999) konnten mit Hilfe von Computersimulationen zeigen, dass hohe Kommunalitäten geringere Stichprobenumfänge ermöglichen und geringe Kommunalitäten höhere Stichprobenumfänge erfordern. Sie geben dem Leser außerdem eine Reihe von Empfehlungen für den Mindeststichprobenumfang in Abhängigkeit von Kommunalitäten und Faktorenzahl mit.

Weiterhin gibt es Anforderungen an die Korrelationsmatrix. So können viele kleine Korrelationen darauf hindeuten, dass keine ausgeprägte Faktorenstruktur im Datensatz existiert (Backhaus et al., 2013, S. 266). Um die Eignung der Korrelationsmatrix konkret zu überprüfen, wurden mehrere Verfahren entwickelt. Backhaus et al. (2013, S. 266 ff.) zählen mehrere davon auf und diskutieren diese: Die inverse Korrelationsmatrix, die Anti-Image-Matrix, das Measure of Sampling Adequacy (MSA), das Kaiser-Meyer-Olkin-Kriterium (KMO) und Bartletts Test auf Spherizität.

In der Literatur gelten das MSA- und KMO-Kriterium als die besten Verfahren, um die Eignung der Korrelationsmatrix für eine Faktorenanalyse zu untersuchen (Backhaus et al. 2013, S. 269), weshalb sie in dieser Arbeit verwendet werden. Sie werden wie folgt berechnet (Klinke 2015):

$$KMO = \frac{\sum\sum_{j \neq k} r_{jk}^2}{\sum\sum_{j \neq k} r_{jk}^2 + \sum\sum_{j \neq k} p_{jk}^2}$$

$$MSA_j = \frac{\sum_{k \neq j} r_{jk}^2}{\sum_{k \neq j} r_{jk}^2 + \sum_{k \neq j} p_{jk}^2}$$

Die Variable r_{jk} ist die Korrelation zwischen den Variablen *j* und *k* und die Variable p_{jk} ist die partielle Korrelation zwischen den Variablen *j* und *k* mit *j,k=1, ... ,p* (Zahl der Variablen). Die partielle Korrelation wird folgendermaßen ausgerechnet: Man berechnet jeweils eine Regression zwischen *j* bzw. *k* (als abhängige Variablen) und allen anderen Variablen des Datensatzes (als unabhängige Variablen). Aus den beiden Regressionen ermittelt man die Residuen und errechnet aus diesen den Korrelationskoeffizienten. Wirken die Faktoren stark auf die Variablen, gehen die partiellen Korrelationen gegen null und die MSA- und KMO-Werte gegen eins. Das MSA-Kriterium ermittelt die Eignung einer einzelnen Variablen für die Faktorenanalyse, während das KMO-Kriterium die Eignung des gesamten Datensatzes ermittelt. Backhaus et al. (2013, S. 269) tragen Interpretationen für unterschiedliche MSA- und KMO-Werte aus der Literatur zusammen: Werte unter 0.5 deuten darauf hin, dass die Daten für eine Faktorenanalyse ungeeignet sind. 0.6 gilt als mittelmäßig und Werte größer als 0.7 als ziemlich gut bis sehr gut (>0.9).

Schritt 2: Faktorenextraktionsmethode(n) auswählen. Die Faktorladungen Λ kann man in der Praxis aus der Kovarianz- und Korrelationsmatrix der Daten extrahieren (Härdle und Simar 2003, S. 283). Um das umsetzen zu können, wurden mehrere Verfahren, sogenannte Faktorenextraktionsmethoden, entwickelt. Brown et al. (2012) listen sieben dieser Verfahren auf: *Principal component analysis* (PCA), *Unweighted least squares* (ULS), *generalized least squares* (GLS), *maximum likelihood* (ML), *principal axis factoring* (PAF), *alpha factoring* (AF) und *image factoring* (IF). In der Fachwelt wird seit langem untersucht und diskutiert, welche davon unter welchen Umständen verwendet werden sollten (Costello und Osborn 2005; Smith und Zygmont 2014). Die am meisten benutzten Extraktionsmethoden sind PCA und PAF (Brown et al. 2012). Costello und Osborn (2005) argumentieren, dass bei multivariater Normalverteilung die ML-Schätzung zu bevorzugen ist und bei multivariater Nichtnormalverteilung PAF

gute Schätzungen berechnet. In dieser Arbeit können diese beiden Methoden jedoch nicht verwendet werden, da sie zu Rechenproblemen, sogenannten Heywood-Fällen (Dillon, Kumar und Mulani 1987; Kolenikov und Bollen 2012), führen.

Als alternative Extraktionsmethode wurde deshalb das Verfahren MINRES (für *minimum residuals*) gewählt, welches in der Praxis stabile Ergebnisse liefert und in der Literatur als gut angesehen wird (Smith und Zygmont 2014). Das MINRES Verfahren führt eine Kleinste-Quadrate-Schätzung zwischen den Elementen der Korrelationsmatrix und den Faktorladungen durch, unter der Nebenbedingung, dass die Kommunalitäten der Faktoren kleiner gleich eins sind (Jöreskog 2003).

Schritt 3: Faktoren bestimmen. Es gibt viele Methoden, um die Zahl der zu extrahierenden Faktoren zu bestimmen. Ledesma und Valero-Mora (2007) listen vier ausgewählte davon auf: Das Ellenbogenkriterium nach Cattell (1966), das MAP-Kriterium nach Velicer (1976), das Eigenwertkriterium nach Kaiser (1960) und die Parallelanalyse nach Horn (1965). Das Eigenwertkriterium nach Kaiser gilt in der Literatur als unpräzise und neigt dazu, zu viele Faktoren auszuwählen (Ledesma und Valero-Mora 2007). Das Ellenbogenkriterium ist subjektiv und kann zu Debatten über die Zahl der zu wählenden Faktoren führen (Brown et al. 2012). Das MAP-Kriterium wird in der Literatur als eine gute Methode angesehen, unterschätzt aber in diversen Situationen die Zahl der Faktoren (Ledesma und Valero-Mora 2007). Schlussendlich wird in der Literatur die Parallelanalyse nach Horn als eine der besten Methoden zur Faktorenextraktion genannt (Costello und Osborne 2005; Smith und Zygmont 2014; Ledesma und Valero-Mora 2007), weswegen sie in dieser Arbeit verwendet wird. Kabacoff (2003, S. 1) beschreibt den Ablauf der Parallelanalyse nach Horn wie folgt:

1. Berechne aus dem Datensatz die Korrelationsmatrix und aus dieser die Eigenwerte. Ordne diese Eigenwerte vom größten Eigenwert (erste Eigenwertposition) zum kleinsten Eigenwert (letzte Eigenwertposition).
2. Erstelle einen zufallsgenerierten Datensatz mit der gleichen Anzahl an Beobachtungen und Variablen wie der originale Datensatz. Die Variablen des zufallsgenerierten Datensatzes sind jeweils voneinander statistisch unabhängig und univariat normalverteilt. Berechne aus dem zufallsgenerierten Datensatz die Korrelationsmatrix und die Eigenwerte von dieser, welche vom größten (erste Eigenwertposition) zum kleinsten (letzte Eigenwertposition) geordnet werden.
3. Wiederhole Schritt zwei insgesamt B mal (z.B. $B=1000$).

4. Berechne aus der Verteilung der *B* Werte jeder Eigenwertposition das 95%-Perzentil. Horn (1965) hat ursprünglich den Mittelwert verwendet, dadurch wird jedoch tendenziell die Zahl der zu extrahierenden Faktoren überschätzt, was mit dem 95%-Perzentil weniger oft passiert (Glorfeld 1995; O'Connor 2000).
5. Vergleiche für jede der Eigenwertpositionen die Eigenwerte der Korrelationsmatrix des Datensatzes mit dem 95%-Perzentil der Eigenwerte der *B* zufallsgenerierten Datensätze. Ist der Eigenwert der Korrelationsmatrix auf einer Eigenwertposition größer als der des 95%-Perzentils der Eigenwerte der *B* zufallsgenerierten Datensätze, wird diese Eigenwertposition (der Faktor) in der Analyse verwendet.

Schritt 4: Rotationsmethode auswählen. In den Gleichungen (1) und (2) gibt es für die Faktorladungen keine eindeutige Lösung (Härdle und Simar 2003, S. 280), sondern (theoretisch) unendlich viele. Aus diesen Lösungsmöglichkeiten gilt es eine auszuwählen, die einfach zu interpretieren ist (Härdle und Simar 2003, S. 280). Doch was ist eine „einfach interpretierbare" Lösung der Faktorladungen? Kieffer (1998) gibt einen Überblick, was in der Literatur darunter verstanden wird:

1. Jede Variable sollte mindestens auf einen Faktor mit null laden.
2. Jeder Faktor sollte von mehreren unabhängigen Variablen mit null geladen werden.
3. Wenn man zwei Faktoren betrachtet, sollten mehrere Variablen auf den einen Faktor mit null laden, aber hoch auf den anderen.
4. Wenn mehr als vier Faktoren existieren, sollten viele Variablen mit null auf jedes Paar an Faktoren laden.
5. Wenn man zwei Faktoren betrachtet, sollten nur wenige Variablen auf beide Faktoren Ladungen haben, welche größer als null sind.

Um aus den zahlreichen Lösungen eine einfache Struktur der Faktorladungen zu finden, wurden diverse Rotationsmethoden entwickelt, welche man in der Literatur (Kieffer, 1998) in zwei Klassen aufteilt: Orthogonale und oblique Rotationsmethoden. Orthogonale Rotationsmethoden errechnen Faktoren, welche nicht korreliert sind - während oblique Rotationsmethoden Faktoren finden, die korrelieren können (aber nicht müssen). Die Ergebnisse orthogonaler Rotationsverfahren sind in der Regel einfacher zu interpretieren als die Ergebnisse von obliquen, da Faktorkorrelationen bei der Interpretation nicht berücksichtigt werden müssen (die Korrelationen sind bei obliquen Verfahren per Definition bei null). Darüber hinaus führen orthogonale Rotationsmethoden weniger oft zu einer Überanpassung an die Daten als oblique Rotationsmethoden. Es wird jedoch kritisiert, dass Faktoren in der Praxis korreliert sein können, was orthogonale Rotationsmethoden jedoch per Verfahren nicht zulassen (Kieffer, 1998). In der Literatur (Kieffer, 1998) wird deshalb vorgeschlagen, sowohl orthogonal als auch oblique zu rotieren,

die Lösungen zu vergleichen und das Rotationsverfahren zu wählen, welches am besten mit der den Daten zugrundeliegenden Theorie einhergeht. Falls bei der obliquen Rotation weitestgehend unkorrellierte Faktoren errechnet werden, wird empfohlen, die orthogonale Lösung zu wählen, da die Korrelationen zwischen den Faktoren sowieso nahe null liegen. Dieser Argumentation wird in dieser Arbeit gefolgt, und die Faktorladungen werden jeweils orthogonal und oblique rotiert.

Es gibt jeweils mehrere Verfahren für die oblique und orthogonale Rotation. Oblique Rotationsmethoden kommen im Allgemeinen zu ähnlichen Ergebnissen (Costello und Osborne 2005, S. 3). In dieser Arbeit wurde das Rotationsverfahren Oblimin gewählt, da es im Gegensatz zu anderen Verfahren bei der Berechnung keine Probleme (z.B. Heywood-Fälle bei der Promaxrotation) verursacht hat. Als Orthogonale Rotationsmethode wird Varimax, die am häufigsten benutzte orthogonale Rotationsmethode (Costello und Osborne 2005, S. 3), verwendet.

Die Idee der Varimaxrotation (Härdle und Simar 2003, S. 290) ist, dass man eine Rotationsmatrix aufstellt, welche durch Multiplikation die Matrix der Faktorladungen abhängig vom Eingabewinkel dreht. Es muss lediglich genau der Eingabewinkel gefunden werden, der die Matrix der Faktorladungen so dreht, dass die Faktorladungen eine einfach zu interpretierende Faktorstruktur haben. Mathematisch wird das über eine Maximierung des Varimaxkriteriums (Härdle und Simar 2003, S. 290) erreicht.

Bei der Obliminrotation werden – vereinfacht gesagt - die Winkel der Faktoren bewusst verändert, sodass sich wiederum die Faktorladungen ändern, bis ein Optimierungskriterium erreicht wird, welches dafür sorgt, dass die Faktorladungen im Sinne der „einfach interpretierbaren Faktorladungsstruktur“ ausfallen (Gorsuch 2014, S. 199 ff.).

Schritt 5: Benennung und Interpretation der Faktoren. Man benennt Faktoren basierend darauf, welche Variablen welchen Faktoren zuzuschreiben sind - und Variablen mit hohen Faktorladungen auf einen Faktor spielen dabei eine besonders große Rolle, da sie den Faktor besonders stark beeinflussen (Brown et al. 2012, S. 6). In der Literatur werden Faktorladungen zwischen 0.3 (Costello und Osborn 2005) und 0.4 (Stevens 2012, S. 333) als Minimum angesehen, damit man eine Variable bei der Interpretation eines Faktors berücksichtigen sollte. Damit ladende Variablen in dieser Arbeit berücksichtigt werden, wird ein Mindestkriterium von >0.4 an die Faktorladungen gestellt.

3.2.4 Bewertung des Faktorenmodells

Anhand der Höhe und Verteilung der Faktorenladungen, dem kumulierten erklärten Varianzanteil des Faktorenmodells und der Interpretierbarkeit der Faktoren kann man beurteilen, wie gut und sinnvoll das Ergebnis einer explorativen Faktorenanalyse ist. Faktorladungen >0.5 werden

als hoch angesehen (Backhaus et al. 2013, Seite 292). Mindestens zwei bis drei Variablen sollten auf einen Faktor laden, damit eine sinnvolle Interpretation des Faktors möglich ist - und fünf mit >0.5 ladende Variablen werden als Indikator für einen guten Faktor gesehen (Brown et al. 2012 und Osborn und Costello, 2005). Mehrere Variablen sollten nicht auf einen Faktor laden, da das die Faktorinterpretation schwierig, bis hin zu unmöglich, macht (Backhaus et al. 2013, S. 292*).* Nach Peterson (2000) wird in der Literatur ein erklärter kumulierter Varianzanteil des Faktorenmodells von 30% bis 40% als schlecht angesehen, 50% wird von manchen als Minimum angesehen, und 90% wird von manchen als gut angesehen. Den kumulierten erklärten Varianzanteil berechnet man, indem man die Summe der Kommunalitäten (die durch die Faktoren erklärte Varianz) des ausgewählten Faktorenmodells durch die Gesamtvarianz teilt.

3.3 Hauptkomponentenanalyse

3.3.1 Ziel und methodische Einordnung

Das Ziel der Hauptkomponentennalyse ist eine Reduktion der Variablenzahl (Härdle und Simar 2003, S.234 ff.). Das wird erreicht, indem - bildlich gesprochen - das ursprüngliche Koordinatensystem rotiert wird, bis ein neues Koordinatensystem (bestehend aus den orthogonalen Hauptkomponenten) gefunden wird, welches die Varianz der Daten bestmöglich abbildet (Klinke, Mihoci und Härdle 2010).

Die Hauptkomponentenanalyse ist verwandt mit der (explorativen) Faktorenanalyse, mit der sie mehrere Eigenschaften teilt (Klinke et al. 2010). Beide Methoden reduzieren zum Beispiel die Zahl der Variablen, beide Methoden nutzen zum Teil ähnliche Rechenverfahren – und beide Methoden können in einigen Fällen zu ähnlichen - oder den gleichen - Ergebnissen kommen (Suhr 2005). Es gibt jedoch auch mehrere Unterschiede zwischen den beiden Methoden: Die Hauptkomponentenanalyse ist ein deskriptives Verfahren, während die (explorative) Faktorenanalyse ein auf umfangreichen Annahmen (Vgl. Abschnitt 3.2.2) basierendes statistisches Modell ist, welches latente Faktoren extrahiert. Es gibt zum Teil erhebliche Unterschiede in Rechenschritten, mathematischen Eigenschaften und praktischen Anwendungsmöglichkeiten (für detaillierte Diskussionen vgl. Costello und Osborn 2005; Klinke et al. 2010; Härdle und Simar 2003, S. 292 ff.; Suhr 2005; Jolliffe 2002).

3.3.2 Funktionsweise der Hauptkomponentenanlyse

Die Idee der Hauptkomponentenalyse ist es, aus einem Koordinatensystem mit p Variablen ein neues Koordinatensystems (bestehend aus orthogonalen Hauptkomponenten als Achsen) mit q ($q<p$) Variablen zu berechnen (Klinke et al. 2010). Es gibt mehrere Verfahren, dieses Ziel zu

erreichen, von denen das Verfahren *best fitting lines* (Klinke 2015) vorgestellt werden soll. Dabei werden die q Hauptkomponenten $PC_j=b_j+a_1t_{j1}$ gesucht, die dieses Optimierungskriterium erfüllen:

$$\underset{(a_j,b_j)}{min} \sum_{i=1}^{n} ||(a_j + b_j t_{ij}) - x_j||^2$$

Die Variable x_i ist dabei die *i-te* (von n) Beobachtungen eines Datensatzes mit p Variablen. Der p-dimensionale Vektor a_j ist der Anstieg der Hauptkomponente PC_j, der p-dimensionale Vektor b_j ist der Achsenabschnitt der Hauptkomponente PC_j und die Variable t_{ij} steht für den Wert der *i-ten* Beobachtung auf der Hauptkomponente PC_j. Um eine eindeutige Lösung zu finden, wird an den Vektor a_j die Bedingung gestellt, dass sein Betrag eins ist und an den Vektor b_j wird die Bedingung gestellt, dass dieser so klein wie möglich sein sollte (Klinke 2015).
Im ersten Schritt des *best fitting lines* Verfahrens wird die erste Hauptkomponente PC_1 iterativ berechnet. Anschließend rechnet man die zweite Hauptkomponente nach demselben Prinzip aus, wobei diese orthogonal auf der ersten liegen muss. Diese Prozedur wird nun für insgesamt q Hauptkomponenten wiederholt.
Damit eine Hauptkomponentenanalyse durchgeführt werden kann, müssen Beziehungen zwischen den Variablen linear sein (Shlens 2014), da die Hauptkomponenten lineare Gleichungen sind, welche nur lineare Variablenzusammenhänge adäquat erfassen können.

3.3.3 Die Hauptkomponentenanalyse in der Praxis

Ähnlich wie die explorative Faktorenanalyse (Abschnitt 3.2.2) wird die Hauptkomponentenanalyse in dieser Arbeit schrittweise durchgeführt, um nachvollziehbar und systematisch vorzugehen:

1. Anforderungen an die Daten überprüfen
2. Hauptkomponenten berechnen und Zahl der Hauptkomponenten auswählen
3. Rotationsmethode auswählen und Hauptkomponenten rotieren
4. Hauptkomponenten interpretieren und benennen

Schritt 1: Anforderung an die Daten überprüfen. Die Variablen müssen lineare Zusammenhänge haben, da nur diese von den Hauptkomponenten erfasst werden können. Weiterhin muss man die Wertebereiche der Variablen im Auge behalten: Unterschiedlich große Wertebereiche der p Variablen zu unterschiedlichen Varianzen führen, obwohl im Verhältnis betrachtet eine gleiche Streuung herrscht. Die Varianzen wiederum beeinflussen die Hauptkomponenten (Shlens 2014). Um dieses Problem beim *best fitting lines* (und anderen Verfahren) Verfahren zu umgehen, werden die Daten deshalb in der Regel standardisiert (Härdle und Simar 2003, S.

233) – z.B. z-standardisiert, sodass jede Variable den Mittelwert null und die Varianz eins hat.

Schritt 2: Hauptkomponenten berechnen und auswählen. Die Hauptkomponenten können über ein Verfahren, wie *best fitting lines*, geschätzt werden. Aus den berechneten Hauptkomponenten muss eine sinnvolle Anzahl ausgewählt werden, wobei die minimale Anzahl an Hauptkomponenten ausgewählt werden sollte, welche die maximale Varianz der Daten erklären (Lorenzo-Seva 2013). Um die Zahl der Hauptkomponenten auszuwählen, wurden viele Verfahren entwickelt (Klinke 2015), von denen sich die Parallelanalyse nach Horn (vgl. Abschnitt 3.2.2) – wie schon bei der Faktorenanalyse – besonders bewährt hat (Franklin, Gibson, Robertson, Pohlmann und Fralish 1995) und deswegen in dieser Arbeit verwendet wird.

Schritt 3: Rotationsmethode auswählen. Hauptkomponenten können rotiert werden (Jolliffe 2002). Wie bei einer Faktorenanalyse (vgl. Abschnitt 3.2.2) kann eine Rotation die Interpretation der Hauptkomponenten verbessern (Richman 1986) - und wie bei der Faktorenanalyse kann man orthogonale und oblique Rotationsmethoden verwenden. Wie bei der explorativen Faktorenanalyse wird jeweils eine Varimax- und Obliminrotation durchgeführt, wobei man anmerken muss, dass die Hauptkomponenten nach einer Obliminrotation nicht mehr orthogonal zueinander sein müssen.

Schritt 4: Hauptkomponenten interpretieren und benennen. Die Ladungen der Variablen auf eine Hauptkomponente geben an, wie stark der Einfluss dieser Variablen auf diese Hauptkomponente ist (Härdle und Simar 2003, S. 241). Die Bennenung der Hauptkomponenten erfolgt entsprechend dieser Zusammenhänge.

Zur besseren Interpretation werden in der Praxis kleine Ladungen von Variablen auf Hauptkomponenten nicht beachtet, wobei hier Faustregeln zwischen >0.5 und >0.8 verbreitet sind (Franklin, Gibson, Robertson, Pohlmann, Fralish 1995) (in dieser Arbeit wird >0.5 verwendet). Wie viel Varianz des ursprünglichen Datensatzes durch die Hauptkomponenten erklärt wird, kann als Kennzahl verwendet werden, wie gut (und aussagekräftig) das Ergebnis einer Hauptkomponentenanalysee ist (Härdle und Simar 2003, S. 241).

3.4 Clustering

3.4.1 Ziel und methodische Einordnung

Ziel des Clusterings ist es, Gruppen von ähnlichen Beobachtungen (Cluster) zu finden, wobei Ähnlichkeit mit Hilfe einer Ähnlichkeitsfunktion, zum Beispiel einem Distanzmaß, erfasst wird (Aggarwal und Zhai 2012). Es gibt eine sehr große Zahl an verschiedenen Clusteringmethoden (Estivill-Castro 2002), die in der Literatur zum Teil unterschiedlich kategorisiert werden (z.B. Tan, Steinbach und Kumar 2005, S.492-495; Soni und Ganatra, 2012, *S.66;* Berkhin 2006;

Rockach und Maimon 2005, S. 330). Von den vielen Clustering-prinzipien soll an dieser Stelle lediglich auf diejenigen eingegangen werden, die in dieser Arbeit verwendet werden: Hierarchisches Clustering, partitionierendes Clustering und dichtebasiertes Clustering (Tan et al. 2005, S. 492 ff.):

1. Bei hierarchischem Clustering geht man davon aus, dass Cluster ineinander hierarchisch verschachtelt sind, z.B zwei Cluster zusammen ein übergeordnetes drittes Cluster bilden können. Hierarchische Clusterverfahren werden in agglomerative und divisive hierarchische Clusterverfahren unterteilt. Beim agglomerativen Clustering behandelt man am Anfang jeden Punkt als einzelnes Cluster und vereinigt bei jedem Schritt die sich am naheliegendsten Cluster zu einem neuen Cluster, bis nach einer bestimmten Anzahl an Schritten nur noch ein Cluster – der komplette Datensatz - übrig bleibt. Beim divisiven Clustering wird die hierarchische Clusterstruktur erarbeitet, indem man mit einem einzelnen Cluster - dem kompletten Datensatz - beginnt und diesen solange aufteilt, bis jeder Punkt ein einzelnes Cluster ist.
2. Bei partitionierenden Clusteringverfahren geht man davon aus, dass die Cluster vollständig voneinander getrennt sind und keine hierarchische Strukutr existiert. Das Ziel partitionierender Verfahren ist es, die bestmögliche Auftrennung der Daten in getrennte Cluster zu finden.
3. Dichtebasierende Clusteringverfahren folgen der Idee, dass Cluster (a) Regionen im Raum sind, in dem hohe Dichten an Punkten herrschen und (b) Cluster durch Bereiche im Raum getrennt sind, in denen (erheblich) niedrigere Dichten herrschen. Was genau ist dabei „Dichte“? Für diesen Begriff gibt es mehrere Definitionen, wobei in dieser Arbeit der zentrumbasierte Dichtebegriff (Tan et al. 2005, S. 528 ff.) verwendet wird (Vgl. Abschnitt 3.4.3). Die zentrumbasierte Dichte wird dabei um eine Beobachtung *b* ermittelt, indem man die Beobachtungen zählt, die sich innerhalb einer bestimmten Distanz um *b* befinden. Je mehr Beobachtungen sich innerhalb einer bestimmten Distanz der Beobachtung *b* befinden, umso höher ist die Dichte.

In dieser Arbeit werden drei Clusterverfahren angewendet: Agglomeratives hierarchisches Clustering, K-means Clustering (ein partitionierendes Verfahren) und OPTICS (ein dichtebasierendes Verfahren). Die Verfahren wurden ausgewählt (Vgl. Abschnitt 3.4.7), da sie unterschiedliche Funktionsweisen, Ansätze und unterschiedliche Stärken und Schwächen haben – und somit die Daten aus verschiedenen Perspektiven untersucht werden können.

3.4.2 Hierarchisches agglomeratives Clustering

3.4.2.1 Distanzen, Ablauf und Verfahren

Es gibt mehrere agglomerative hierarchische Clusteringalgorithmen, die oftmals ein ähnliches Grundprinzip haben (Murtagh 1983, S.237 ff.): Man berechnet den Abstand zwischen allen Beobachtungen und vereinigt genau die Beobachtungen zu Clustern, die sich jeweils am nächsten liegen. Danach werden schrittweise immer genau die Cluster miteinander vereinigt, die den geringsten Abstand zueinander haben. Nach einer bestimmten Zahl an Schritten ist nur noch ein großes Cluster (der komplette Datensatz) übrig und die hierarchische Clusterstruktur wurde herausgearbeitet. Damit man dieses Clusterprinzip durchführen kann, benötigt man Kennzahlen, um (a) den Abstand zwischen Beobachtungen und (b) den Abstand zwischen Clustern zu berechnen (Backhaus et al. 2013, S. 331 ff.):

Abstandsberechnung zwischen Beobachtungen. Um den Abstand zwischen Beobachtungen zu messen, wurden für unterschiedlich skalierte Variablen unterschiedliche Abstandsmaße entwickelt (Backhaus et al. 2013, S. 331 ff.). Da im Datensatz dieser Arbeit nur metrisch skalierte Variablen vorkommen, wird an dieser Stelle lediglich das Abstandsmaß, welches im Allgemeinen (Härdle und Simar 2003, S. 305) für dieses Skalenniveau benutzt wird - die euklidische Distanz - vorgestellt:

$$d(x, y) = \|x - y\|_2$$

Die Variablen x und y stehen hierbei für die Beobachtungen mit p Variablen und sind als p-dimensionale Vektoren formuliert.

Abstandsberechnung zwischen Clustern. Um den Abstand zwischen Clustern zu messen und zu entscheiden, welche Cluster vereinigt werden sollen, gibt es mehrere sogenannte Linkage-Verfahren (Tan et al. 2005, S. 517 ff.). Härdle und Simar (2003, S. 309 ff.) beschreiben unter anderem die Linkageverfahren Single Linkage, Complete Linkage, Average Linkage, Centroid Linkage und Wards Methode:

Single Linkage. Bei diesem Verfahren ermittelt man die Distanz zwischen zwei Clustern, indem man die Distanz (zum Beispiel die euklidische Distanz) der beiden voneinander am kürzesten entfernten Beobachtungen der zwei Cluster ermittelt.

Complete Linkage. Bei diesem Verfahren ermittelt man die Distanz zwischen zwei Clustern, indem man die Distanz der beiden voneinander am weitesten entfernten Beobachtungen der zwei Cluster ermittelt.

Average Linkage. Bei dieser Methode berechnet man für jede Beobachtung des einen Clusters alle Abstände zu jeder Beobachtung des anderen Clusters und bildet daraus den Durchschnitt.

Centroid Linkage. Bei dieser Methode ist die Clusterdistanz zwischen zwei Clustern der Abstand zwischen ihren Clusterzentren, den sogenannten Clustercentroiden (z.B. der Mittelwert aller Punkte eines Clusters).

Wards Methode. Bei Wards Methode werden genau zwei Cluster miteinander vereinigt, wenn die Varianz nach ihrer Vereinigung im Vergleich zu den anderen Vereinigungsmöglichkeiten so gering wie möglich steigt.

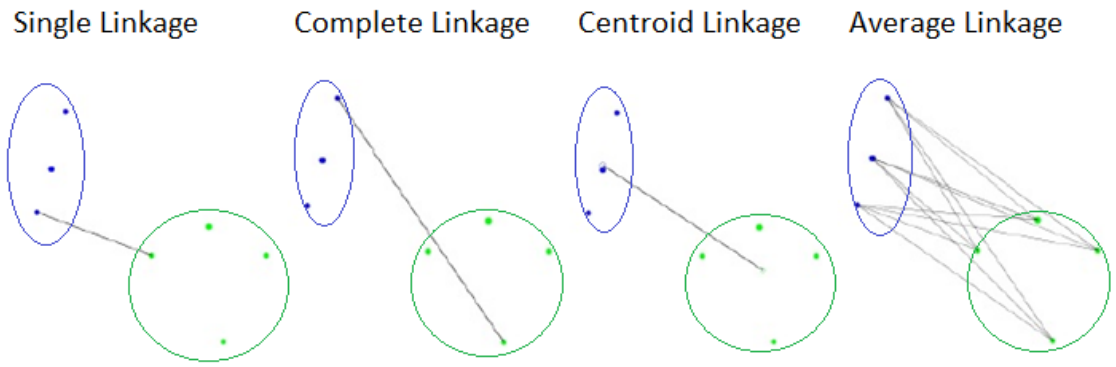

Abbildung 1*: Veranschaulichung der Distanzenermittlung für vier verschiedene Linkageverfahren. Quelle: Klinke 2015, S. 108 ff.; anschließend bearbeitet.*

Den Abstand zwischen zwei Clustern kann man in jedem Vereinigungsschritt durch die Formel nach Lance und Williams berechnen (Härdle und Simar 2003, S. 309). Diese geht davon aus, dass zwei Cluster *P* und *Q* vereinigt werden und man den Abstand aus diesem neuen Cluster *P* + *Q* zu einem Cluster *R* feststellen möchte:

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|.$$

Der Ausdruck *d(R, P+Q)* steht dabei für die Distanz zwischen dem Cluster *R* und dem neu erschaffenem Cluster aus *P* und *Q*. Die Werte δ_1 bis δ_4 sind je nach Linkageverfahren unterschiedlich und können Tabelle 4 entnommen werden.

Welchen dieser Linkageverfahren sollte man verwenden? Punj und Stewart (1983, S. 138 ff.) untersuchten zwölf empirische Studien, in denen verschiedene Linkageverfahren gegeneinander getestet wurden. Wards Methode schnitt dabei - neben dem Average Linkage Verfahren – als eines der besten Verfahren ab und wird deshalb im praktischen Teil dieser Arbeit verwendet.

Name	δ_1	δ_2	δ_3	δ_4
Single linkage	1/2	1/2	0	-1/2
Complete linkage	1/2	1/2	0	1/2
Average linkage	1/2	1/2	0	0
Centroid	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0

Tabelle 4: Werte für δ_1 bis δ_4 für unterschiedliche Linkageverfahren. Quelle: Härdle und Simar 2003, S. 309; anschließend bearbeitet.

3.4.2.2 Bestimmung der Clusterzahl

Das agglomerative hierarchische Clustering ermittelt die hierarchische Clusterstruktur. Die Herausforderung ist nun, dass man die Zahl der Cluster basierend auf der hierarchischen Clusterstruktur festlegen muss. Um das zu machen, wurde eine große Zahl an Grafiken, Verfahren und Kennzahlen entwickelt (Tan et al. 2005; Jung, Park, Du und Drake 2003; Milligan und Cooper 1985; Tibshirani, Walther und Hastie 2001; Sugar und James 2011). Zwei davon sollen vorgestellt werden: Die grafische Bestimmung der Clusterzahl mit einem Dendrogramm und die rechnerische Bestimmung der Clusterzahl mit dem Silhouettenkoeffizient nach Roussew.

Das Dendrogram wurde ausgewählt, da es die hierarchische Clusterstruktur gut veranschaulichen kann (Härdle und Simar 2003, S. 310). Der Silhouettenkoeffizient wurde ausgewählt, da er eine in der Praxis oft benutzte (Tan et al. 2005, S. 541), intuitive Kennzahl ist und in empirischen Tests gegenüber ähnlichen Kennzahlen sehr gut abschneidet (Handel, Knowles, Kell 2005; Liu, Li, Xiong, Gao und Wu 2010).

Dendrogramm. Ein Dendrogramm ist eine Grafik (Vgl. Abbildung 2 und 3), welche die hierarchische Clusterstruktur visualisieren kann. Auf der x-Achse sind die Beobachtungen gruppiert nach Clustern angeordnet und auf der y-Achse sind die Distanzen (bei Wards Methode die Varianz) abgetragen (Härdle und Simar 2003, S. 310). Die Clusterstruktur wird durch einen Baum dargestellt, der immer in zwei Teile aufgesplittet wird. Große Veränderungen des Baums, gemessen als Abstand auf der y-Achse, zeigen, dass eine Vereinigung zwischen vergleichsweise weit voneinander entfernten Clustern erfolgt.

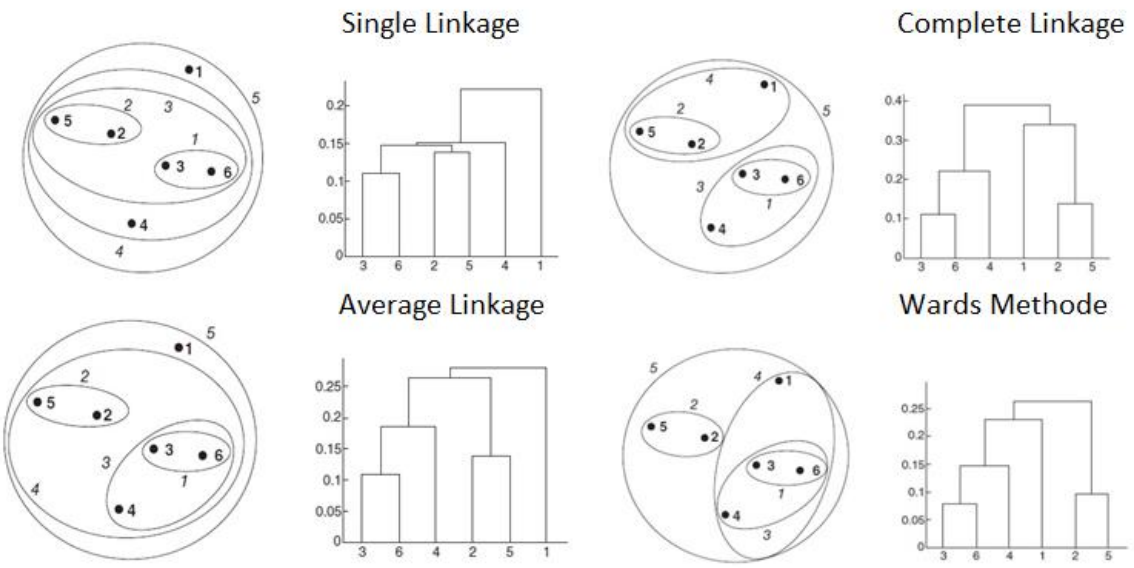

Abbildung 2: Dendrogramme für verschiedene Linkageverfahren. Die Daten sind in allen Fällen gleich. Das Linkageverfahren hat Einfluss auf die Clusterdistanzen – und damit auf die Clusterwahl. Quelle: Tan et al. 2005, S. 520-522; anschließend bearbeitet.

Mit diesen Informationen kann man die Clusterzahl wie folgt bestimmen: Man wählt, von oben ausgehend, genau den Punkt nach einer Aufspaltung mit vergleichsweise großer Veränderung in der Distanz (Vgl. Abbildung 3).

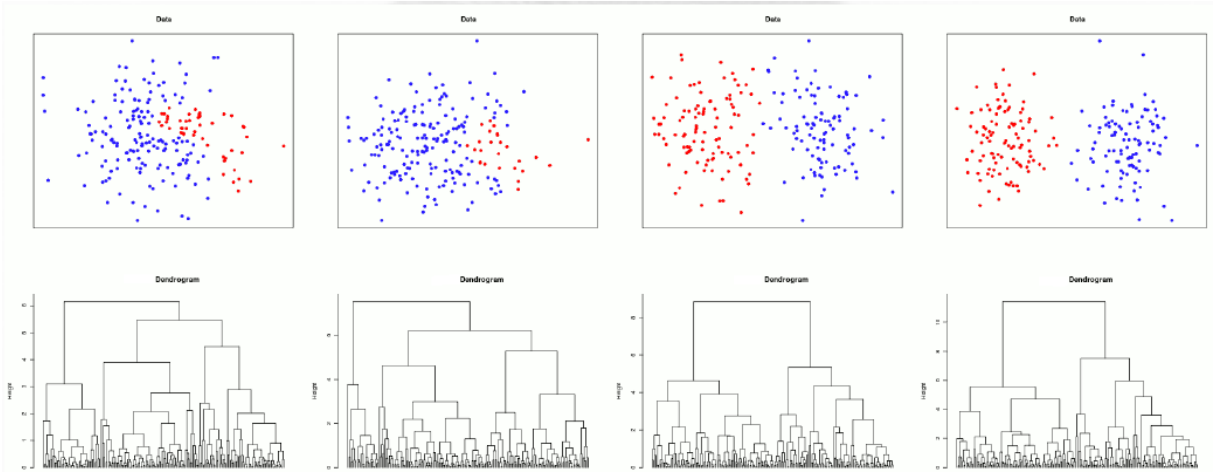

Abbildung 3: Dendrogramme für vier unterschiedliche Datensätze. Je besser die Cluster getrennt sind, umso größer werden die Höhenunterschiede (=Distanzen) in den Bäumen im Dendrogramm. Quelle: Klinke 2015, S. 129; anschließend bearbeitet.

Der Silhouttenkoeffizient. Der Silhouettenkoeffizient *s(i)* für eine Beobachtung *i*, welche einem Cluster zugeordnet worden ist, wurde von Rousseeuw (1987) wie folgt definiert:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}.$$

Die Variable *i=1, ... ,n* steht für eine Beobachtung des Datensatzes und *a(i)* ist der Durchschnitt der Distanzen einer Beobachtung *i* eines Clusters *A* zu allen anderen Beobachtungen im Cluster *A*. Die Variable *b(i)* ist der Durchschnitt der Distanzen einer Beobachtung im Cluster *A* zu allen Beobachtungen des Clusters *B*. Das Cluster *B* ist dabei das Cluster, welches die kürzeste Distanz zum Cluster *A* hat. Ein kleines *a(i)* deutet darauf hin, dass zwischen der Beobachtung *i* und allen anderen Beobachtungen im Cluster A im Schnitt eine geringe Distanz herrscht. Ein großes *b(i)* deutet darauf hin, dass die Beobachtung *i* des Clusters *A* zu allen Beobachtungen im Cluster *B* im Schnitt eine große Distanz hat. Je kleiner *a(i)* und je größer *b(i)*, umso mehr geht *s(i)* gegen den Maximalwert eins. Der Minimalwert für *s(i)* ist minus eins.
Um die gesamte Clusterlösung zu beurteilen, kann man den Mittelwert aller *s(i)* errechnen. Ein hoher Wert (bei eins) bedeutet, dass die Beobachtungen in einem Cluster relativ eng liegen und dass die Cluster räumlich relativ weit und gut voneinander getrennt sind, während Werte um null darauf hindeuten können, dass keine Clusterstruktur existiert (Rousseeuw 1987).

3.4.3 OPTICS

3.4.3.1 Definitionen

OPTICS (Ankerst, Breunig, Kriegel und Sander, 1999), für „ordering points to identify cluster structure“, ist ein dichtebasierter Clusteringalgorithmus. Er ist eine Weiterentwicklung von dem Algorithmus DBSCAN (Ester, Kriegel, Sander und Xu 1996). Um den dichtebasierten Ansatz des OPTICS-Algorithmus beschreiben zu können, wurden von Ankerst et al. (1999) mehrere Begriffe eingeführt und definiert:

Die Distanz *ε*. Die Variable ε ist die Bezeichnung für einen konkreten Wert einer beliebigen Distanzfunktion (in dieser Arbeit der euklidischen Distanz) und muss vom Benutzer festgelegt werden.

Die *ε*-Umgebung. Die ε-Umgebung ist eine Menge an Beobachtungen, die innerhalb der Distanz ε um eine Beobachtung x_i liegt. Eine Beobachtung x_j ($i,j=1, ...,n$ und $i \neq j$) gehört zur ε-Umgebung von x_i, wenn die Distanz zwischen x_i und x_j kleiner als der Wert ε ist.

Die minimale Punktanzahl *MinPts*. *MinPts* ist die Menge an Beobachtungen, die in der ε-Umgebung um eine Beobachtung x_i liegen muss, damit diese als Kernbeobachtung gilt. Die Beobachtung x_i zählt sich selbst zu *MinPts*. *MinPts* ist einer der beiden Eingabeparameter und muss vom Benutzer festgelegt werden.

Die Kernbeobachtung. Sind in der ε-Umgebung von einer Beobachtung x_i mindestens *MinPts* Beobachtungen, gilt diese als Kernbeobachtung.

Direkte Dichterreichbarkeit. Eine Beobachtung x_i ist direkt dichteerreichbar von einer Beobachtung x_j, wenn x_i zur ε-Umgebung von x_j gehört und x_j eine Kernbeobachtung ist.

Dichteereichbarkeit. Eine Beobachtung x_i ist dichteerreichbar von einer Beobachtung x_j, wenn es eine Beobachtungskette $x_1, \ldots, x_n$ mit $x_1 = x_i$ *und* $x_n = x_j$ gibt, in der x_{i+1} jeweils direkt dichteereichbar von x_i ist.

Dichteverbundenheit. Eine Beobachtung x_i ist mit einer Beobachtung x_j dichteverbunden, wenn es eine Beobachtung o gibt, von dem aus beide Beobachtungen dichterreichbar sind. Die Beobachtung o kann x_i oder x_j selbst sein.

Cluster und Rauschen. Ein Cluster ist eine Menge von Beobachtungen des Datensatzes. Damit Beobachtungen ein Cluster bilden, müssen folgende Bedingungen erfüllt sein:

1. Eine Kette von zueinander dichteverbundenen Beobachtungen bildet ein Cluster.
2. Wenn x_i zu einem Cluster gehört und x_j von x_i aus dichteerreichbar ist, gehört x_j ebenfalls zu einem Cluster.

Beobachtungen im Datensatz, die keinem Cluster zugeordnet sind, gelten als Rauschen.

Die Kerndistanz. Ist eine Beobachtung x_i eine Kernbeobachtung, so ist ihre Kerndistanz das kleinste ε, das eine Beobachtung durch ihre ε-Umgebung gerade noch zu einer Kernbeobachtung machen würde. Hat eine Beobachtung eine geringe Kerndistanz, bedeutet das, dass die Dichte um die Beobachtung hoch ist. Eine große Kerndistanz bedeutet dagegen, dass die Dichte um eine Beobachtung gering ist.

Ist eine Beobachtung keine Kernbeobachtung, wird ihre Kerndistanz auf „nicht definiert“ gesetzt. Wird ε auf einen Wert gesetzt, der größer ist als die Distanz zwischen den beiden voneinander entferntesten Beobachtungen, tritt dieser Fall niemals ein, da jede Beobachtung eine Kernbeobachtung ist.

Die Erreichbarkeitsdistanz. Falls die Beobachtung o eine Kernbeobachtung ist, ist die Erreichbarkeitsdistanz zwischen den Beobachtungen o und x_i durch das Maximum aus der Kerndistanz von o und der Distanz zwischen o und x_i gegeben. Ist o keine Kernbeobachtung, wird die Erreichbarkeitsdistanz auf „nicht definiert“ gesetzt. Wird ε auf einen Wert gesetzt, der größer ist als die Distanz zwischen den beiden voneinander entferntesten Beobachtungen, tritt dieser Fall niemals ein, da jede Beobachtung eine Kernbeobachtung ist.

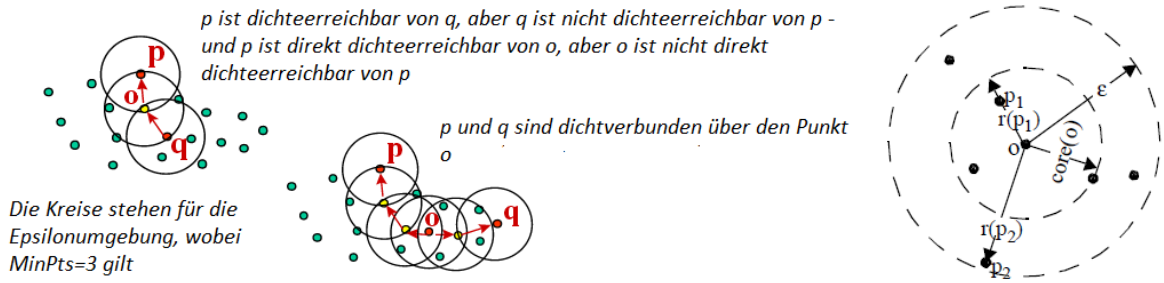

Abbildung 4: Begriffsdarstellung von OPTICS. Die Variable *core(o)* steht für die Kerndistanz von *o* (für *MinPts=4*), $r(p_1)$ und $r(p_2)$ stehen für die Erreichbarkeitsdistanzen zu den Punkten p_1 und p_2. Quelle: Ankerst et al. (1999, S. 3 ff.); anschließend bearbeitet.

3.4.3.2 Funktionsweise

Das Funktionsprinzip des OPTICS-Algorithmus lässt sich vereinfacht wie folgt darstellen:

1. Gib ε und *MinPts* ein.
2. Wähle die erste Beobachtung des Datensatzes aus und wiederhole Schritt drei und vier solange, bis alle Beobachtungen des Datensatzes als „besucht“ markiert sind.
3. Berechne für eine Beobachtung x_i die ε-Umgebung, die Kerndistanz und die Erreichbarkeitsdistanz zu allen anderen Beobachtungen in ihrer ε-Umgebung. Speichere diese Werte und markiere die Beobachtung x_i als „besucht“.
4. Sortiere die Beobachtungen der Epsilonumgebung von x_i anhand ihrer Erreichbarkeitsdistanz (von der kleinsten zur größten). In der dadurch gewonnenen Reihenfolge werden nun diese Beobachtungen (die ε-Umgebung von x_i) nach Schritt 3 abgearbeitet, wenn sie noch nicht als „besucht“ markiert wurden.
5. Wenn alle Beobachtungen des Datensatz als „besucht“ markiert sind, gib am Ende die Reihenfolge der Abarbeitung, die Kerndistanzen und die Erreichbarkeitsdistanzen (jeweils zur nächstliegenden Beobachtung) von allen Beobachtungen an.

Der OPTICS-Algorithmus berechnet *keine* Cluster. Stattdessen sortiert er die Beobachtungen in der Reihenfolge ihrer Abarbeitung und speichert deren Kerndistanzen und Erreichbarkeitsdistanzen. Durch diese Informationen hat man die Clusterstruktur des Datensatzes erfasst (und kann die Cluster herausarbeiten), welche durch einen Erreichbarkeitsplot dargestellt werden kann.

Auf der x-Achse enthält dieser die Beobachtungen in der Reihenfolge ihrer Abarbeitung (was bedeutet, dass jeweils Beobachtungen mit kleiner Erreichbarkeitsdistanz im Erreichbarkeitsplot nah beieinander liegen). Auf der y-Achse ist die Erreichbarkeitsdistanz zur jeweils am geringsten entfernten Beobachtung abgetragen.

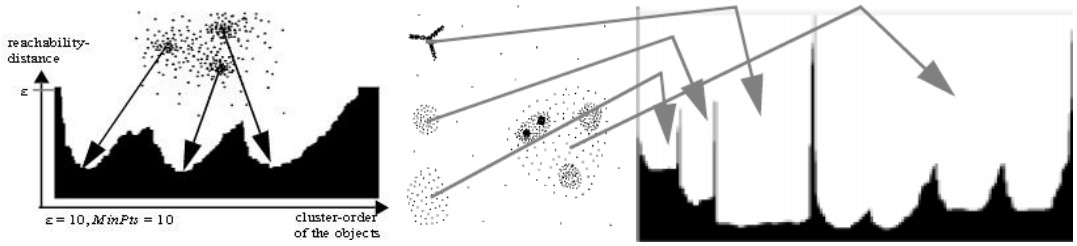

Abbildung 5: Erreichbarkeitsplots für unterschiedliche zweidimensionale Datensätze. Wie man im rechten Bild sehen kann, werden verschachtelte Cluster mit unterschiedlichen Dichten gut erkannt. Quelle: Ankerst et al. (1999, S.6 ff.); anschließend bearbeitet.

3.4.3.3 Clusterextraktion

Grafisch betrachtet ist ein Cluster eine Senkung in einem Erreichbarkeitsplot. Je dichter die Beobachtungen in einem Cluster beeinander liegen, umso tiefer ist die Senkung. Zusätzlich zu den Clustern kann man auch die hierarchische Clusterstruktur mit Hilfe der geordneten Erreichbarkeitsdistanzen (Ankerst et al. 1999, S.9 ff.) aus dem Erreichbarkeitsplot herausarbeiten. Das Prinzip dahinter: Clusterhierarchien führen dazu, dass es – bildlich gesprochen - „Senkungen in den Senkungen“ im Erreichbarkeitsplot gibt (Vgl. Abbildung 4 und 5).

Rechnerisch lassen sich Cluster und Clusterhierarchien durch folgendes Prinzip bestimmen: Immer wenn ein Cluster beginnt oder endet, fällt oder steigt die Erreichbarkeitsdistanz über einige wenige Beobachtungen um einen vergleichsweise großen Prozentsatz (Vergleich dazu die roten Markierungen im Erreichbarkeitsplot aus Abbildung 5) – während mitten in einem Cluster keine starken Änderungen der Erreichbarkeitsdistanzen zu finden sind.

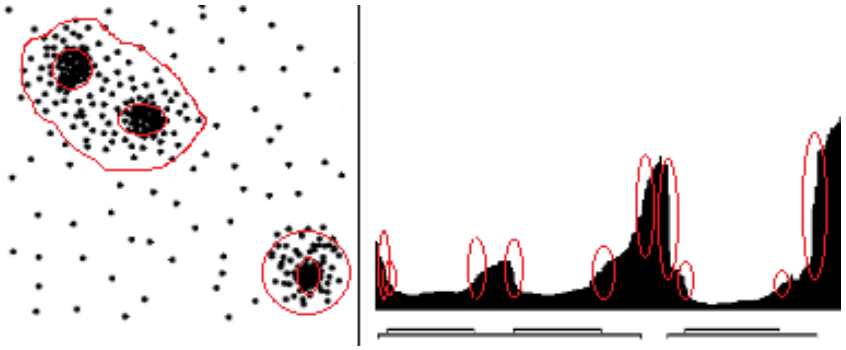

Abbildung 6: Herausarbeitung von hierarchischen Clusterstrukturen mit Hilfe des Erreichbarkeitsplots. Die roten Umkreisungen markieren die ξ-Steigungsgebiete. Quelle: Ankerst et al. (1999, S. 9); bearbeitet.

Diese Änderungen der Erreichbarkeitsdistanzen der geordneten Beobachtungen wird als ξ-Veränderung bezeichnet (Ankerst et al. 1999, S.9). Die ξ-Veränderung gibt an, dass sich die Er-

reichbarkeitsdistanz im Erreichbarkeitsplot von einer Beobachtung zur nächsten um einen Betrag von ξ% verändert. In der Nähe der Clusterzentren sind die ξ-Veränderungen in der Regel nahe null, aber an den Clusterrändern nehmen die ξ-Veränderungen deutlich größere Werte an. Liegen mehrere Punkte mit vergleichsweise hohen ξ-Veränderungen im Erreichbarkeitsplot nebeneinander, bezeichnet man dieses Gebiet als ξ-Steigungsgebiet (Vgl. dazu Abbildung 5). Ein Cluster findet man nun, indem man seine beiden ξ-Steigungsgebiete (eins am Anfang und eins am Ende) ermittelt. Sind Cluster ineinander verschachtelt, liegt das untergeordnete Cluster zwischen den zwei ξ-Steigungsgebieten des übergeordneten Clusters - und die ξ-Steigungsgebiete des untergeordneten Clusters sind von den ξ-Steigungsgebieten des übergeordneten Clusters durch Bereiche, welche keine ξ-Steigungsgebiete sind, getrennt.

3.4.3.4 Parameterwahl

Die Parameter *MinPts* und ε müssen vom Benutzer festgelegt werden und haben Einfluss auf die Rechengeschwindigkeit und die Clusterlösung:

Epsilon. Je kleiner man ε wählt, umso kleiner wird die abzusuchende ε-Umgebung einer jeden Beobachtung. Damit müssen bei kleinen ε-Werten Erreichbarkeitsdistanzen und Kerndistanzen zu vergleichsweise wenigen Beobachtungen ermittelt werden (was den Rechenprozess beschleunigt). Wählt man jedoch ein zu kleines ε, können Cluster mit geringer Dichte nicht mehr geordnet abgesucht und erkannt werden. Im Erreichbarkeitsplot erkennt man das daran, dass alle Erreichbarkeitsdistanzen oberhalb von ε nicht abgetragen werden (vgl. dazu den Wert UNDEF (für nicht definiert) in Abbildung 4).

Setzt man ε auf einen Wert, der größer ist als die Distanz zwischen den zwei voneinander am weitesten entfernten Punkten im Datensatz, ist die Epsilonumgebung eines jeden Punktes der komplette Datensatz. Dadurch ist die Rechenkomplexität entsprechend hoch, jedoch geht man kein Risiko ein, Cluster mit geringer Dichte fälschlicherweise zu übersehen.

In dieser Arbeit wird ε wie folgt festgesetzt: Zuerst werden experimentell große Werte für ε ($\varepsilon=100$ für z-standartisierten Datensatz, wenn nötig höher) ausprobiert, sodass die Erreichbarkeitsdistanzen zu allen Beobachtungen kalkuliert werden können. Danach wird ε schrittweise solange heruntergesetzt wird, bis man die Clusterstruktur ohne erheblichen Informationsverlust im Erreichbarkeitsplot gut erkennen kann.

MinPts. Wird *MinPts* auf einen kleinen Wert gesetzt, können kleine Gruppierungen von Ausreißern in den Daten (fälschlicherweise) als Cluster erkannt werden. Wird *MinPts* auf einen zu großen Wert gesetzt, können Cluster mit wenigen Beobachtungen nicht erkannt werden. In der Literatur findet man nur wenige (nicht begründete) Faustregeln für die Wahl von *MinPts*.

Manchmal wird *MinPts=2p-1* (Zhang, 2006) oder *MinPts=p+1* (Hennig 2015) festgelegt, wobei *p* für die Anzahl der Variablen im Datensatz steht.

Ankerst et al. (1999, S. 5 und 6) merken an, dass die Clusterstruktur insgesamt nur schwach auf moderate Änderungen von *MinPts* reagiert, sodass man von einem Wertebereich für *MinPts* ausgehen kann, bei welchem OPTICS die hierarchische Clusterstruktur angemessen herausarbeitet. Um diesen Wertebereich zu finden, wird *MinPts* in dieser Arbeit zuerst experimentell auf Werte zwischen *p* und *3p* gesetzt, sodass man einen Überlick über die hierarchische Clusterstruktur in Abhängigkeit von *MinPts* bekommt. Nachdem ein grober Wertebereich für *MinPts* bestimmt wurde, wird *MinPts* in diesem Wertebereich weiter experimentell geändert, bis ein vergleichsweise stetiger Verlauf der Erreichbarkeitsdistanzen im Erreichbarkeitsplot erreicht wird (vgl. Abbildung 7).

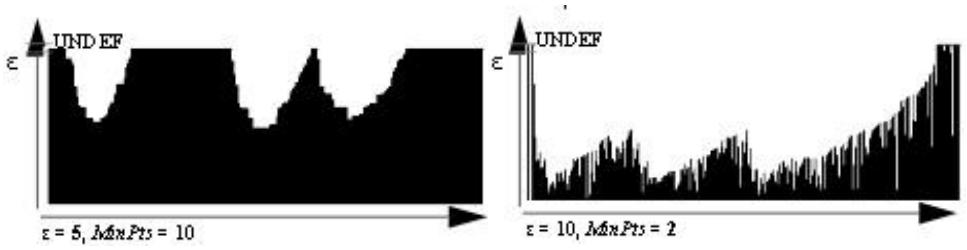

Abbildung 7: Einfluss der Eingabeparameter auf die geordneten Erreichbarkeitsdistanzen. Die Grundstruktur der Cluster wird nicht wesentlich durch die Eingabeparameter verändert. Quelle: Ankerst et al. (1999); bearbeitet.

3.4.4 K-means

3.4.4.1 Entwicklung und Methodik

K-means ist ein Clusteringverfahren, dessen Grundprinzip von einer Reihe von Wissenschaftlern in den 1950er und 1960er zum Teil unabhängig entdeckt wurde (Bock 2007). Die Bezeichnung K-means stammt ursprünglich von MacQueen (1967) (Bock 2007), jedoch existieren mehrere Versionen des K-means Algorithmus, die eine ähnliche Grundidee haben, jedoch in den konkreten Rechenschritten zum Teil unterschiedlich vorgehen (Morisette und Chartiert 2013). Das allgemeine Funktionsprinzip läuft in mehreren Schritten ab (Tan et al. 2005, S. 497 ff.):

1. Wähle *k* zufällig ausgewählte Beobachtungen des Datensatzes als Clusterzentren aus.
2. Ordne die Beobachtungen genau dem Cluster zu, dessen Varianz durch die neue Beobachtung am geringsten erhöht wird. Mathematisch wird damit die Summe der Clustervarianzen (SSE – *„sum of squared error“*) über alle Cluster minimiert:

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

Die Variable *x* ist eine Beobachtung des Datensatzes mit *p* Variablen, C_i ist das *i-te* von *K* Clustern, welche jeweils das Clusterzentrum c_i haben. Je kleiner *SCV* ist, umso näher sind alle Beobachtung an den jeweiligen Clusterzentren und umso besser ist die Clusterlösung.

3. Berechne die *k* Clusterzentren der Cluster neu, wobei das Clusterzentrum c_i der Durchschnitt aller Beobachtungen im Cluster ist.
4. Gehe wieder zu 2., oder stoppe den Algorithmus, wenn sich die Clusterzentren im Vergleich zur letzten Iteration nicht verändert haben oder ein anderes Abbruchkriterium erreicht wurde. Abbruchkriterien in der Praxis sind zum Beispiel eine bestimme Anzahl von Iterationen des Algorithmus oder dass sich ein bestimmter Prozentsatz an Beobachtungen, der von einer Iteration zu anderen das Cluster wechselt, unterschritten wird.

3.4.4.2 Clusterzahl bestimmen

Damit man eine Clusterung mit K-means durchführen kann, muss man die Zahl der Cluster, *k*, kennen (bzw. bestimmen). Zur Bestimmung von *k* wurden zahlreiche Methoden entwickelt, welche in bestimmten Situationen jeweils Stärken und Schwächen haben (Jung et al. 2003; Milligan und Cooper 1985; Tibshirani et al. 2001; Sugar und James 2011; Charrad, Ghazzali, Boiteau, Niknafs, Charrad 2014). Ritter (2014, S. 169) merkt an, dass solche Verfahren zur Bestimmung von *k* nicht immer zu einer eindeutigen Lösung führen.

In dieser Arbeit wird zur Bestimmung von *k* einem Ansatz von Tan et al. (2005, S. 546 ff.) gefolgt, nach welchem zwei Informationen zur Bestimmung von *k* benutzt werden:

1. Der Silhouettenkoeffizient wird für mehrere Clusterlösungen (in dieser Arbeit: *k=2, ...,15*) ermittelt und als Funktion von *k* dargestellt.
2. Die Summe der Clustervarianzen (SCV) wird für mehrere Clusterlösungen (in ideser Arbeit *k=2, ...,15*) ermittelt und als Funktion von k dargestellt.

Die beiden Funktionen werden in einer Grafik abgetragen und das *k* ausgewählt, bei dem der Silhouettenkoeffizient so groß wie möglich und die SCV so klein wie möglich ist. Sollten die beiden Kriterien in dieser Arbeit nicht zum selben Ergebnis kommen, wird zwischen ihnen und der Clusterzahl abgewogen - wobei kleinere *k* im Zweifel bevorzugt werden, um die Modellkomplexität potentiell zu begrenzen.

3.4.5 Clusterbewertung

Viele Clusterverfahren finden oft auch Cluster in Daten, in denen keine Clusterstruktur existiert (Tan et al. 2005, S. 532). Das macht es nötig zu bewerten, wie gut die Lösung eines Clusterverfahrens ist (Clustervalidierung). Es gibt zwei Möglichkeiten, Clusterlösungen zu bewerten (Rendón, Abundez, Arizmendi und Quiroz 2011, S. 1 ff.): Die interne und die externe Bewertung.

Bei der externen Bewertung lässt man Clusterverfahren Cluster in einem Datensatz finden, von denen man die wahre Clusterstruktur bereits kennt. Anschließend werden die berechneten Cluster mit den tatsächlichen Clustern verglichen, sodass die Qualität des Clusterverfahrens festgestellt werden kann. Wenn man jedoch Clusteranalyse als exploratives Verfahren nutzt – was in dieser Arbeit der Fall ist - und man die Clusterlösung nicht kennt, kann man generell keine externe Bewertung machen.

Eine interne Bewertung der Clusterstruktur ist jedoch trotzdem möglich. Die Idee der internen Bewertung ist wie folgt: Eine ausgeprägte (und wünschenswerte) Clusterstruktur bedeutet, dass die Beobachtungen in einem Cluster nah beieinander liegen und die Cluster durch weniger Bereiche getrennt sind, in denen die Beobachtungen weit auseinander liegen. Um diesen Sachverhalt zu messen, wurden eine Reihe von Indizes und Koeffizienten entwickelt und empirisch bewertet (Handel, Knowles und Kell 2005; Rendon, Abundez, Arizmendi und Quiroz 2011; Halkidi, Batistakis und Vazirgiannis 2002; Halkidi und Vazirgiannis 2001; Brock, Pihur, Datta und Datta 2008; Kovács, Legány und Babos 2005; Liu, Li, Xiong, Gao und Wu 2010).

In dieser Arbeit wird der der Silhouettenkoeffizient nach Rousseeuw zur internen Clustervalidierung eingesetzt (wobei er beim agglomerativen hierarchischem Clustering und beim K-means Clustering ja bereits zur Bestimmung der Clusterzahl verwendet wird), da er in Untersuchungen gut abschnitt (Handel et al. 2005; Liu, Li, Xiong, Gao und Wu 2010).

3.4.6 Variablenbeurteilung der Cluster

Hill, Lewicki und Lewicki (2006, S.122) schlagen vor, die Mittelwerte der Cluster für jede einzelne Variable mit Hilfe einer einfaktoriellen ANOVA (*„analysis of variance“*) zu Vergleichen, um herauszufinden, in welchen Variablen sich die Cluster besonders unterscheiden. Ursprünglich ist eine einfaktorielle ANOVA als parametrischer Test konzipiert, welcher die Hypothese überprüft, ob alle Mittelwerte von zwei oder mehr Gruppen auf der untersuchten metrisch skalierten Variable (der Zielvariable) verschieden sind (Lee, Lee und Lee 2000, S. 486). Wie Hill et al. (2006, S. 122) jedoch beispielhaft zeigen, kann man die Teststatistik *f* der

ANOVA als deskriptives Maß verwenden, um herauszufinden, wie verschieden die Gruppenmittelwerte vom Gesamtmittelwert sind. Je kleiner f ist, umso mehr weichen die Mittelwerte der k Gruppen vom Gesamtmittelwert ab.

3.4.7 Datenstandardisierung in der Clusteranalyse

Im euklidischen Raum kann es dann passieren, dass Variablen mit einem großen Wertebereich starken Einfluss auf die Clusterlösung haben, weil auf diesen Variablen (im Vergleich zu Variablen mit kleinem Wertebereich) Cluster allein durch den Wertebereich weiter voneinander entfernt sind (Cooper und Milligan 1988, S. 181 ff.), und dadurch besser getrennt erscheinen. Die Wertebereiche der Variablen können somit die „wahre“ Clusterstruktur verzerren. In der Literatur (Cooper und Milligan 1988, S. 181 ff.) wird u.a. Datenstandardisierung vorgeschlagen, um dieses Problem in den Griff zu bekommen. In dieser Arbeit wird dabei die z-Standardisierung verwendet, sodass jede Variable so transformiert wird, dass sie jeweils den Mittelwert null und die Varianz eins hat (Cooper und Milligan 1988, S. 183). Die Clusteranalyse mit agglomerativen hierarchischen Clustering, K-Means und OPTICS wird auf den standardisierten Daten durchgeführt. Es werden auch nichtstandardisierte Daten geclustert (aber nicht tiefergehend analysiert, z.B. keine Bennenung der Cluster), um den Einfluss des Wertebereichs auf die Clusterung zu analysieren.

3.4.8 Beurteilung der Clusteringalgorithmen

Unterschiedliche Clusterverfahren haben unterschiedliche Vor- und Nachteile. K-means wird zum Beispiel als recheneffizienter Algorithmus bewertet, der konvexe, räumlich gut getrennte Cluster erkennen kann - jedoch auch mehrere Schwachstellen hat (Tan et al. 2005, S. 497 ff.):

1. Ausreißer können starken Einfluss auf die Clusterlösung haben.
2. Nicht-konvexe Cluster, nicht gut getrennte Cluster und Cluster mit unterschiedlichen Dichten können von K-means oftmals nicht erkannt werden.
3. Der Parameter k - der entscheidend für die Clusterlösung sein kann - muss in der Praxis bestimmt werden (oder bekannt sein).
4. Durch die zufällig gewählten Startpunkte kann es bei mehrfacher Clusterung mit K-means beim selben Datensatz zu unterschiedlichen Clusterlösungen kommen.
5. K-means, als partitionierendes Verfahren, erkennt hierarchisch verschachtelte Cluster grundsätzlich nicht.

Agglomeratives hierarchisches Clustering hat gegenüber K-means vor allem den Vorteil, dass es hierarchische Clusterstrukturen herausarbeiten kann (Tan et al. 2008, S. 524 ff.), jedoch hat diese Form des Clusterings ebenfalls mehrere Schwachstellen:

1. Durch das schrittweise herausarbeiten der Clusterstruktur können agglomerative hierarchische Verfahren die lokale Clusterstruktur gut erkennen, sie finden jedoch nicht zwangsläufig die globale Clusterstruktur.
2. Viele der Linkageverfahren haben, wie auch der K-Means-Algorithmus, Probleme beim Erkennen von nicht-konvexen Clustern.

Viele der Schwachstellen vom K-Means Clustering und agglomerativen hierarchischen Clustering hat OPTICS nicht (Deepak und Roy 2006; Kriegel et al. 1999): OPTICS erkennt die unterschiedlichsten Clusterformen. Auch komplex ineinander verschachtelte Cluster mit sehr unterschiedlichen Dichten können erkannt werden. OPTICS erkennt Ausreißer sehr gut (und diese können anschließend als solche klassifiziert werden), während bei K-Means und agglomerativem hierarchischem Clustering Ausreißer in bestimmte Cluster zugeordnet werden (was die Clusterlösung verfälschen kann). Mithilfe von OPTICS kann man die hierarchische Clusterstruktur herausarbeiten (im Gegensatz zu K-means).

Allerdings hat auch OPTICS Schwachstellen: Die Parameter *MinPts* und ε müssen in der Praxis, oftmals experimentell (vgl. Abschnitt 3.4.3.3), bestimmt werden. Da OPTICS ein vergleichsweise rechenintensives Verfahren sein kann (wenn man ε zu groß wählt), kann die mehrfache Durchführung von OPTICS bei sehr großen Datensätzen zeitaufwändig werden.

4 Datenanalyse

4.1 Deskriptive Statistiken

Um einen ersten Überblick über die Daten zu bekommen, wird für alle Variablen jeweils das Minimum (*Min*), das erste Quartil (Q_1), der Median (Q_2), das dritte Quartil (Q_3) und das Maximum (*Max*) errechnet:

	Min	Q_1	Q_2	Mw	Q_3	Max
num_self_hrefs	0.00	1.00	2.00	2.80	4.00	56.00
num_videos	0.00	0.00	0.00	0.64	0.00	75.00
num_imgs	0.00	1.00	1.00	1.81	1.00	51.00
num_hrefs	0.00	4.00	7.00	9.36	12.00	122.00
n_non_stop_unique_tokens	0.00	0.65	0.70	0.70	0.76	0.97
n_unique_tokens	0.00	0.48	0.55	0.55	0.61	0.87
n_tokens_content	0.00	244.00	400.00	539.87	727.00	6336.00
average_token_length	0.00	4.53	4.69	4.69	4.86	6.38
n_non_stop_words	0.00	1.00	1.00	1.00	1.00	1.00
n_tokens_title	3.00	9.00	10.00	10.28	12.00	19.00
abs_title_subjectivity	0.00	0.17	0.50	0.34	0.50	0.50
title_sentiment_polarity	-1.00	0.00	0.00	0.08	0.14	1.00
title_subjectivity	0.00	0.00	0.07	0.25	0.46	1.00
abs_title_sentiment_polarity	0.00	0.00	0.00	0.14	0.21	1.00
LDA_00	0.10	0.51	0.70	0.66	0.84	0.92
LDA_01	0.02	0.03	0.04	0.08	0.05	0.71
LDA_02	0.02	0.03	0.04	0.08	0.05	0.80
LDA_03	0.02	0.03	0.03	0.07	0.05	0.84
LDA_04	0.02	0.03	0.04	0.12	0.16	0.82
global_sentiment_polarity	-0.24	0.09	0.14	0.14	0.19	0.62
rate_positive_words	0.00	0.67	0.75	0.74	0.83	1.00
global_subjectivity	0.00	0.39	0.44	0.44	0.49	1.00
min_positive_polarity	0.00	0.03	0.10	0.09	0.10	0.70
global_rate_negative_words	0.00	0.01	0.01	0.01	0.02	0.06
avg_negative_polarity	-1.00	-0.30	-0.24	-0.24	-0.18	0.00
global_rate_positive_words	0.00	0.03	0.04	0.04	0.05	0.12
rate_negative_words	0.00	0.17	0.25	0.26	0.33	1.00
min_negative_polarity	-1.00	-0.70	-0.50	-0.48	-0.25	0.00
max_positive_polarity	0.00	0.60	0.80	0.77	1.00	1.00
max_negative_polarity	-1.00	-0.12	-0.10	-0.11	-0.05	0.00
avg_positive_polarity	0.00	0.31	0.35	0.35	0.40	0.80

Tabelle 5: Zusammenfassung der Häufigkeitsverteilung der untersuchten Variablen. Quelle: Eigene Darstellung.

Die Hälfte aller Artikel enthält weniger als acht Links zu verschiedenen Websiten. Mindestens 75% der Artikel haben (mindestens) ein Bild und weniger als 25% der Artikel haben Videos. Insgesamt gibt es nur vergleichsweise wenige Artikel, welche sehr viele Links, sehr viele Videos und sehr viele Bilder enthalten.

Die Hälfte aller Artikel enthält 400 Tokens oder weniger, aber es gibt einige Artikel, welche weitaus mehr als 400 Tokens enthalten, was man daran erkennt, dass der Mittelwert (539 Tokens) viel größer als der Median (400 Tokens) ist. Überschriften enthalten drei bis 19 Tokens, wobei der Mittelwert und Median bei zehn Tokens liegt.

Die erfassten Themen sind auf die Artikel sehr ungleich verteilt: Während die Hälfte aller Artikel eine Nähe >0.7 (von maximal eins) zum Thema *LDA_00* haben (und 25% eine Nähe >0.84 zu diesem Thema haben), haben weniger als 25% aller Artikel eine Nähe >0.05 zu den Themen *LDA_01, LDA_02, LDA_03* und weniger als 25% aller Artikel haben eine Näher >0.16 zu dem Thema LDA_04. Das bedeutet, dass die Themen *LDA_01 – LDA_04* nur für einen sehr kleinen Anteil der Wirtschaftsnachrichten eine Rolle spielen, während Thema *LDA_00* die Wirtschaftsnachrichten eindeutig dominiert.
Die Subjektivität der Überschriften ist bei der Hälfte der Artikel null und nur mehr als 25% der Artikel haben eine Subjektivität >0.21. Mehr als 75% aller Überschriften hat eine neutrale oder positive Sentimentalität. Auch die Texte enthalten überwiegend positive Wörter und vergleichsweise wenige Wörter.
Eine Überraschung gab es bei der Analyse der Variable *n_non_stop_words*: Eigentlich soll diese Variable den Anteil der Wörter im Text angeben, welche keine Stoppwörter sind. Der Wertebereich müsste folglich zwischen null und eins liegen – aber es gibt *nur* die Ausprägungen null und (approximativ) eins. Inhaltlich gesehen ist das nicht schlüssig. Nach mehrfachem herunterladen der Daten (Fernandes et al. 2015b) bin ich zu dem Schluss gekommen, dass der Fehler direkt im Datensatz liegen muss und die Kennzahl vermutlich falsch kalkuliert wurde. Sie wurde deshalb vor der weiteren Analyse aus dem Datensatz entfernt.

4.2 Korrelationsanalyse

Die Korrelationen wurden in zwei Schritten untersucht. Im ersten Schritt wurde die Korrelationsmatrix erzeugt, wobei die Korrelationskoeffizienten zur besseren Übersucht durch eine Farbskala ersetzt wurden (vgl. Abbildung 8). Darüber hinaus wurden die Variablen so angeordnet, dass jeweils die Variablen, welche vergleichsweise stark miteinander korrelieren, nebeneinander liegen. Im zweiten Schritt wurde alle paarweisen Plots – bei 30 Variablen sind das 465 Stück – einzeln untersucht, um nichtlineare Strukturen (Cluster, nichtlineare funktioniale Zusammenhänge, …) zu entdecken, die der Korrelationskoeffizienten nicht angemessen abbilden kann. Dabei wurden vereinzelt nichtlineare bzw. und nichtlinear funktionale Zusammenhänge entdeckt (vgl. Abbildung 9).
Es wurde die Entscheidung getroffen, die Variablen *abs_title_sentiment polarity* und *abs_title_subjectivity* von der Hauptkomponenten-, Faktoren- und Clusteranalyse auszuschließen. Die beiden Variablen sind Betragsfunktionen von den Variablen *title_sentiment polarity* bzw. *title_subjectivity*, welche durch den Korrelationskoeffizienten nicht angemessen dargestellt werden können (was jedoch wichtig für die Hauptkomponenten- und Faktorenanalyse ist).

In der Clusteranalyse können diese Variablen durch das Funktionsprinzip der Betragsfunktion genau entgegengesetzte Werte zu den Variablen *title_subjectivity* und *title_sentiment_polarity* erzeugen, was (möglicherweise unerwünschten) Einfluss auf die Clusterstruktur haben kann.

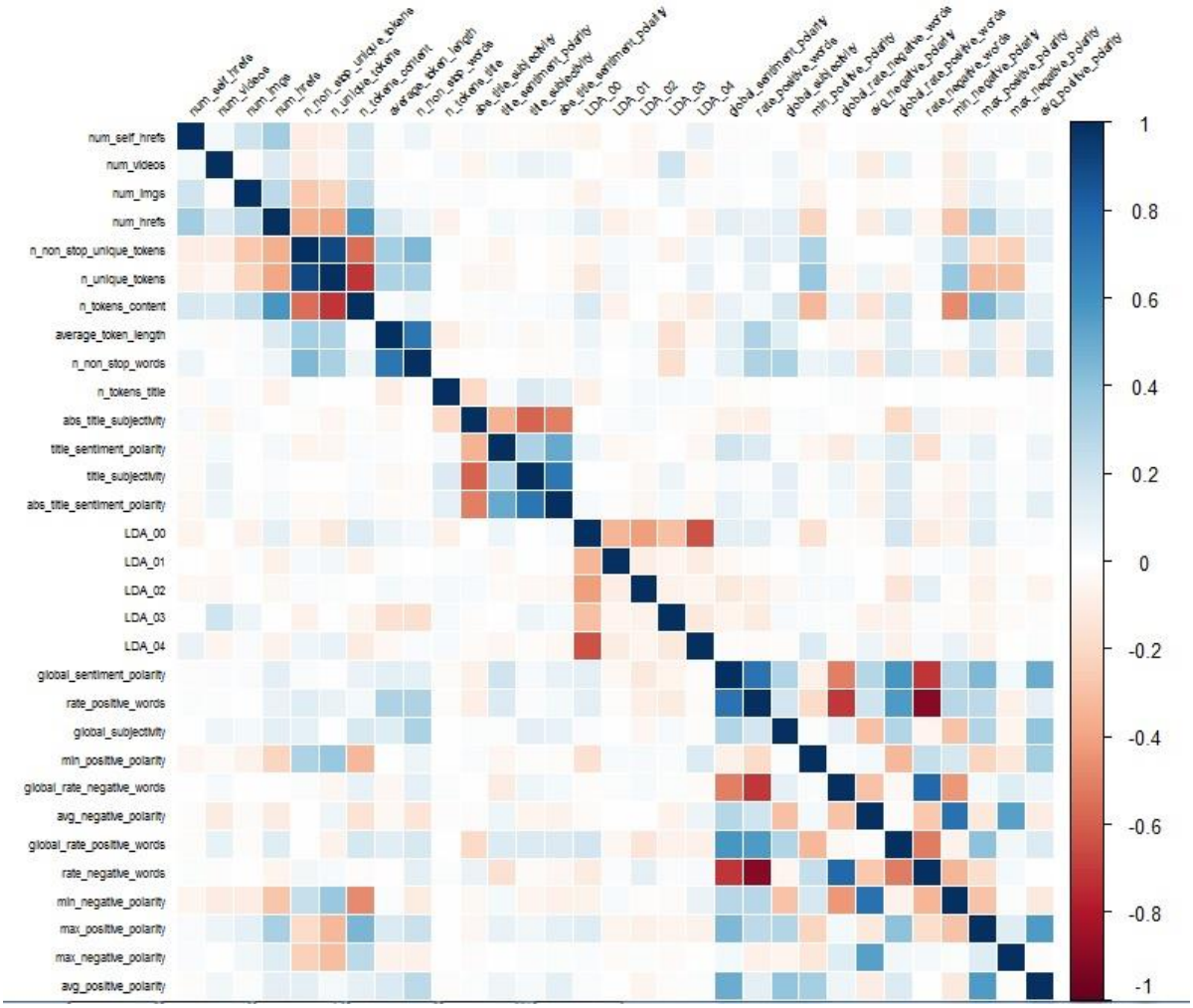

Abbildung 8: Die Korrelationsstruktur der Daten. Die spezielle Anordnung der Variablen zeigt, dass die Korrelationen in Blöcken (einige Variablen korrelieren miteinander, aber kaum mit den restlichen Variablen) zusammengefasst werden können. Quelle: Eigene Darstellung.

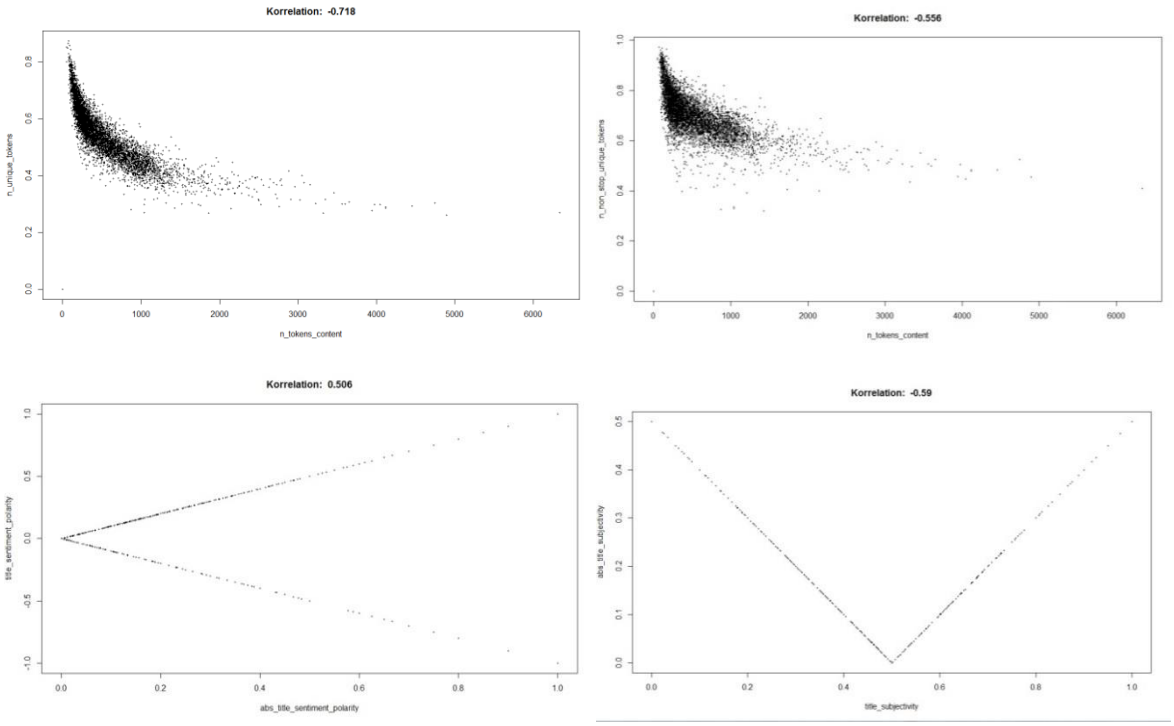

Abbildung 9: Nichtlineare Zusammenhänge zwischen Variablen des Datensatzes. Quelle: Eigene Darstellung.

4.3 Explorative Faktorenanalyse

4.3.1 Vorbereitung der explorativen Faktorenanalyse

Vor der Durchführung der Faktorenanalyse wurde die Unabhängigkeit der Beobachtungen, das Skalenniveau der Variablen, der Stichprobenumfang und die Anforderungen an die Korrelationsmatrix untersucht.

Unabhängigkeit der Beobachtungen: Artikel von Mashable (die Beobachtungen) verlinken oft (vgl. Abschnitt 4.1) zu anderen Artikeln von dieser Website, weshalb zwischen vielen Artikeln eine zeitliche und (potentiell) eine inhaltliche Abhängigkeit bestehen kann. Fernandes et al. (2015a und b) sprechen diese (mögliche) Abhängigkeit nicht an und haben auch keine Kennzahlen erfasst, um diese Abhängigkeit untersuchen zu können. An dieser Stelle kann deswegen nur darauf hingewiesen werden, dass eine zeitliche und inhaltliche Abhängigkeit zwischen den Artikeln bestehen *kann*. Es kann aber nicht gesagt werden, ob und in welchem Maße sie tatsächlich vorkommt. Sollten die Artikel (die Beobachtungen) inhaltlich in großem Maße abhängig sein, können die Ergebnisse der Analyseverfahren in dieser Arbeit ihre Gültigkeit verlieren.

Skalenniveau und Stichprobenumfang. Alle Variablen sind metrisch skaliert. Es gibt 6258 Beobachtungen und 28 untersuchte Variablen, was ein Beobachtung-zu-Variablen Verhältnis von 223.5:1 ergibt. Der Stichprobenumfang ist damit auch für schwach ausgeprägte Faktorenstruktur geeignet.

Analyse der Korrelationsmatrix. Die Faktorenanalyse verlangt lineare Zusammenhänge zwischen den Variablen. Einige Variablen haben nichtlineare Strukturen (vgl. Abschnitt 4.2), was bereits zum Ausschluss von zwei Variablen geführt hat.

Neben (einigen wenigen weiteren) nichtlinearen Strukturen verursacht die Korrelationsmatrix darüber hinaus noch ein weiteres Problem: Mit den verbliebenen 28 Variablen ist sie singulär, weshalb bei der Berechnung der Hauptkomponenten, Faktoren und MSA-Werte mithilfe des Softwarepaketes R Probleme auftraten.

Durch einen Ausschluss mehrerer niedrig korrelierender Variablen konnte eine Korrelationsmatrix gefunden werden, welche nicht singulär ist. Ausgeschlossen wurden die Variablen *LDA_01, LDA_02, LDA_03, LDA_04*. Diese Variablen korrelierten in nennenswerten Ausmaß (vgl. Abschnitt 4.2) nur mit der Variable LDA_00.

Das KMO-Kriterium (0.65) und die MSA-Werte (vgl. Tabelle 6) zeigen, dass der Datensatz für eine Faktorenanalyse grundsätzlich geeignet ist, wenn auch nicht besonders gut (vgl. Abschnitt 3.2.3).

num_self_hrefs	0.58	global_rate_negative_words	0.67	n_non_stop_unique_tokens	0.62
num_videos	0.42	avg_negative_polarity	0.48	title_sentiment_polarity	0.67
num_imgs	0.70	global_rate_positive_words	0.62	title_subjectivity	0.54
num_hrefs	0.77	rate_negative_words	0.66	average_token_length	0.33
LDA_00	0.81	min_negative_polarity	0.65	global_sentiment_polarity	0.74
n_unique_tokens	0.64	max_positive_polarity	0.80	rate_positive_words	0.70
n_tokens_content	0.77	max_negative_polarity	0.40	global_subjectivity	0.77
n_tokens_title	0.48	avg_positive_polarity	0.49	min_positive_polarity	0.68

Tabelle 6: MSA-Werte der einzelnen Variablen. Quelle: Eigene Darstellung.

Faktorenanzahl bestimmen, Faktoren extrahieren, Faktoren rotieren: Die Parallelanalyse nach Horn (vgl. Abbildung 10) wurde mit *B=250* durchgeführt und ergab, dass acht Faktoren extrahiert werden sollten. Die Ergebnisse der Faktorenextraktion mit MINRES, jeweils einmal mit Varimax- und Obliminrotation, ist in Tabelle 7 zu finden.

Faktoren interpretieren. Die Interpretationen der varimax- und obliminrotiertn Modelle sind Abschnitten 4.3.2 und 4.3.3 beschreiben. Faktorladungen welche keine Beträge >0.4 hatten, wurden – wie in Abschnitt 3.3 beschrieben und begründet – in der Analyse dieser Arbeit nicht beachtet.

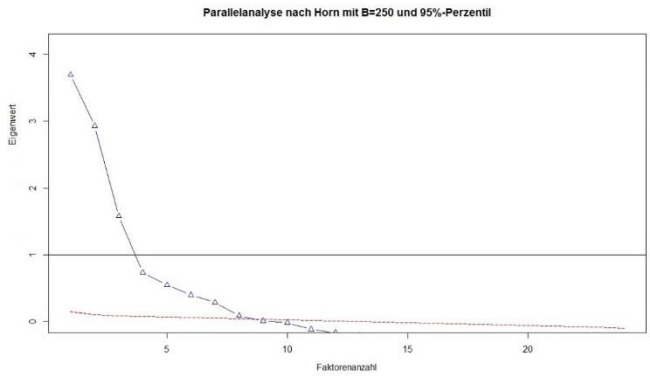

Abbildung 10: Eine Parallelanalyse nach Horn. Die Dreiecke sind die Eigenwerte des Datensatzes, die roten Punkte sind die 95%-Perzentile der zugehörigen Eigenwertpositionen aus *B=250* zufallsgenerierten Datensätzen. Quelle: Eigene Darstellung.

4.3.2 Exploratives Faktorenmodell mit Varimaxrotation

Faktor f_1: Die Variablen *global_sentiment_polarity* und *rate_positive_words* laden positiv auf diesen Faktor. Die Variablen *rate_negative_words* und *global_rate_negative_words* laden negativ auf ihn. Hat ein Artikel einen hohen positiven Wert auf diesem Faktor, bedeutet dass, dass der Anteil positivpolarer Wörter vergleichsweise hoch, und der Anteil negativpolarer Wörter vergleichsweise gering ist. Aus diesem Verhältnis folgt in vielen Fällen eine gewisse Polaritätsstärke mit. Die kompakte Bennennung des Faktors wurde deshalb so gewählt: „Verhältnis der Anteile positivpolarer und negativpolarer Wörter an allen Wörtern, und die daraus resultierender Polarität".

Faktor f_2: Die Variablen *n_unique_tokens* und *non_stop_unique_tokens* laden positiv auf diesen Faktor. Die Variable *n_tokens_content* lädt negativ auf diesen Faktor. Faktor f_2 kann man damit als „Wortschatzdichte" (Wortschatz je Textumfang) interpretieren. Hat ein Artikel einen hohen positiven Wert auf diesem Faktor, kann man erwarten, dass ein großer Wortschatz im Verhältnis zum Textumfang vorliegt.

Faktor f_3: Die Variablen *global_sentiment_polarity*, *global_subjectivity*, *max_positive_polarity* und *avg_positive_polarity* laden positiv auf diesen Faktor. Dieser Faktor lässt sich damit als „Polaritätsstärke positiv polarer Wörter mit resultierender Subjektivität" interpretieren. Auf den ersten Blick ist er Faktor f_1 ähnlich (die Variable *global_sentiment_polarity* lädt auf beide Faktoren). Bei genauerer Analyse erkennt man jedoch den Unterschied: Faktor f_1 gibt den *Anteil* von positivpolaren zu negativpolaren Wörtern im Text an (was zu Polarität im Text führt), während f_3 etwas darüber aussagt, *wie stark* positiv polar (und subjektiv) die positivpolaren Wörter sind. Hat ein Artikel einen hohen Wert auf diesem Faktor, bedeutet das, dass die positiven

Wörter in diesem Artikel vergleichsweise hoch positiv polar sind (unabhängig davon, wie viele positiv polare Wörter in dem Artikel vorkommen).

Faktor f_4: Die Variablen *avg_negative_polarity* und *min_negative_polarity* laden positiv auf diesen Faktor. Dieser Faktor lässt sich als „Polaritätsstärke der negativ polaren Wörter" (unabhängig davon wie viele negative Wörter sind) interpretieren. Hat ein Artikel einen hohen Wert auf diesem Faktor, bedeutetet dass, dass die negativen Wörter im Artikel vergleichsweise hoch negativ polar sind.

Faktor f_5: Die Variable *global_rate_positive_words* ist die einzige, welche mit >0.4 auf diesen Faktor lädt. Der Faktor ist damit äquivalent zu dieser Variable („Anteil positiv polarer Wörter im Text") zu interpretieren.

Faktor f_6: Die Variablen *avg_negative_polarity* und *max_negative_polarity* laden positiv auf diesen Faktor. Dieser Faktor ist damit stark mit f_4 verwandt (*avg_negative_polarity* lädt ebenfalls auf f_4) und ist ähnlich wie dieser zu interpretieren (die Hauptkomponentenealyse – jeweils mit Varimax- und Obliminrotation – findet eine Hauptkomponente, welche eine Vereinigung aus f_4 und f_6 mit ähnlichen Ladungsbeträgen ist, vgl. Tabelle 12)

Faktor f_7: Die Variablen *num_self_hrefs und num_hrefs* laden positiv auf diesen Faktor. Der Faktor lässt sich als „Verlinkungsumfang" interpretieren. Hat ein Artikel einen hohen Wert auf diesem Faktor, bedeutetet das, dass es im Artikel viele Links zu verschiedenen Websiten gibt.

Faktor f_8: Die einzige Variable, die auf diesen Faktor lädt (mit 0.71), ist *avg_token_length*. Der Faktor ist damit äquivalent zu dieser Variable („durchschnittliche Anzahl der Zeichen je Token") zu interpretieren.

	f_1^v	f_2^v	f_3^v	f_4^v	f_5^v	f_6^v	f_7^v	f_8^v	u_v^2	h_v^2	f_1^o	f_2^o	f_3^o	f_4^o	f_5^o	f_6^o	f_7^o	f_8^o	u_o^2	h_o^2
num_self_hrefs							0.43		0.81	0.19						0.51			0.81	0.19
num_videos									0.93	0.07									0.93	0.07
num_imgs									0.84	0.16									0.84	0.16
num_hrefs							0.67		0.42	0.58						0.71			0.42	0.58
n_non_stop_unique_tokens		0.87							0.12	0.88		0.90							0.12	0.88
n_unique_tokens		0.96							0.00	1.00		0.99							0.00	1.00
n_tokens_content		-0.65							0.25	0.75		-0.50							0.25	0.75
average_token_length								0.71	0.37	0.63								0.75	0.37	0.63
n_tokens_title									0.99	0.01									0.99	0.01
title_sentiment_polarity									0.96	0.04									0.96	0.04
title_subjectivity									0.96	0.04									0.96	0.04
LDA_00									0.95	0.05									0.95	0.05
global_sentiment_polarity	0.66		0.54						0.13	0.87	-0.42		0.52						0.13	0.87
rate_positive_words	0.90								0.03	0.97	-0.80								0.03	0.97
global_subjectivity			0.45						0.65	0.35									0.65	0.35
min_positive_polarity									0.58	0.42									0.58	0.42
global_rate_negative_words	-0.88								0.11	0.89	0.97								0.11	0.89
avg_negative_polarity				0.69		0.66			0.00	1.00				0.75			0.49		0.00	1.00
global_rate_positive_words					0.82				0.15	0.85					0.89				0.15	0.85
rate_negative_words	-0.96								0.01	0.99	0.89								0.01	0.99
min_negative_polarity				0.79					0.11	0.89				0.92					0.11	0.89
max_positive_polarity			0.60						0.41	0.59			0.52						0.41	0.59
max_negative_polarity						0.81			0.28	0.72							0.83		0.28	0.72
avg_positive_polarity			0.91						0.16	0.84			0.93						0.16	0.84

Tabelle 7: Faktorladungen für Faktormodelle mit Varimax- und Obliminrotation. Das hochgestellte *v* steht für Varimax, das hochgestellt *o* für Oblimin. Quelle: Eigene Darstellung.

4.3.3 Explorative Faktorenmodell mit Obliminrotation

Wenn man nur Ladungen >0.4 betrachtet (vgl. Abschnitt 3.2.3), sind die Faktoren f_2,f_4,f_5 und f_8 der obliminrotierten Faktorenmodells fast identisch zu den entsprechenden Faktoren des varimaxrotierten Faktorenmodells. Die Ladungsbeträge unterscheiden sich jeweils nur um wenige Prozent (vgl. Tabelle 7). Die Faktoren f_6 und f_7 des varimaxrotierten Faktorenmodells sind jeweils fast identisch mit den Faktoren f_7 und f_6 des obliminrotierten Faktorenmodells.

Die Variablen, welche auf f_1 der Varximaxrotation laden, laden jeweils mit dem entgegengesetzten Vorzeichen (und leicht veränderten Ladungsbeträgen) auf f_1 der Obliminrotation. Damit ist der Faktor f_1 der Obliminrotation - grafisch betrachtet - approximativ der Faktor f_1 der Varimaxrotation in umgekehrte Richtung.

Der Faktor f_3 des obliminrotierten Faktorenmodells ist dem Faktor f_3 des varimaxrotierten Faktorenmodells ähnlich: Die Variablen *global_sentiment_polarity*, *max_positive_polarity* und *avg_positive_polarity* laden positiv auf beide Faktoren - der einzige Unterschied ist, dass die Variable *global_subjectivity* positiv auf den Faktor f_3 der Varimaxrotation lädt, und nicht auf f_3 der Obliminrotation (alle anderen Ladungen sind fast identisch).

Die Interpretationen der varimax- und obliminrotierten Faktoren sind zur besseren Übersicht in Tabelle 8 dargestellt:

Faktor (Varimax)	Faktor (Oblimin)	Faktorinterpretation
1	1	Verhältnis der Anteile positivpolarer und negativpolarer Wörter, inkl. resultierender Polarität
5	5	Anteil positiv polarer Wörter
2	2	Wortschatzdichte
3		Polaritätsstärke positiv polarer Wörter mit resultierender Subjektivität und Polarität
	3	Polaritätsstärke positiv polarer Wörter mit resultierender Polaritätsstärke
4, 6	4, 7	Polaritätsstärke der negativ polaren Wörter
7	6	Verlinkungsumfang
8	8	Durchschnittliche Anzahl der Zeichen je Token

Tabelle 8: Übersicht über die Interpretationen der Faktoren. Quelle: Eigene Darstellung.

Es gibt Korrelationen mit Beträgen >0.3 zwischen den obliminrotierten Faktoren (vgl. Tabelle 9). So ist f_1 (Verhältnis der Anteile positivpolerer und negativpolarer Wörter, inkl. resultierender Polarität) jeweils mit f_4 (Polaritätsstärke der negativ polaren Wörter) und f_5 (Anteil positiv polarer Wörter) negativ korreliert; f_2 (Wortschatzdichte) ist mit f_6 (Verlinkungsumfang) negativ korreliert. Insgesamt kann man sagen, dass viele Faktoren des obliminrotierten Faktorenmodells mit Korrelationsbeträgen zwischen 0.2 und 0.5 miteinander korrellieren.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
f_1	1.00	0.01	-0.13	-0.35	-0.35	-0.05	0.05	-0.11
f_2	0.01	1.00	0.04	0.25	-0.09	-0.50	-0.29	0.21
f_3	-0.13	0.04	1.00	-0.06	0.24	0.13	0.04	0.20
f_4	-0.35	0.25	-0.06	1.00	-0.08	-0.24	0.24	-0.13
f_5	-0.35	-0.09	0.24	-0.08	1.00	0.11	0.12	0.20
f_6	-0.05	-0.50	0.13	-0.24	0.11	1.00	0.17	0.24
f_7	0.05	-0.29	0.04	0.24	0.12	0.17	1.00	0.03
f_8	-0.11	0.21	0.20	-0.13	0.20	0.24	0.03	1.00

Tabelle 9: Die Korrelationen zwischen den acht obliminrotierten Faktoren. Quelle: Eigene Darstellung.

4.3.4 Vergleich der Faktorenmodelle

Beide Rotationsverfahren kommen zu ähnlichen - zum Teil zu fast identischen! - Faktoren mit sehr ähnlichen Faktorladungen. Zu einem gewissen Teil ist dass der Auswahl der Faktorladungen geschuldet: Zieht man zur Interpretation Beispielsweise alle Faktorladungen mit einem Betrag >0.20 zur Analyse hinzu, unterscheiden sich die Lösungen der beiden Rotationsmethoden stärker (sie sind jedoch auch dann noch recht ähnlich). In dieser Arbeit wurden allerdings nur Ladungsbeträge >0.4 verwendet, da in anderen Szenarien sehr viele Variablen gleichzeitig auf mehrere Faktoren laden - und die Faktoreninterpretation dadurch (und die vielen zusätzlichen kleinen Faktorladungen) erheblich verkompliziert werden würde.

4.3.5 Bewertung der Faktorenmodelle

Bei dem Faktorenmodell mit Varimaxrotation haben drei der acht Faktoren mehr als zwei Variablen, welche mit >0.4 auf die Faktoren laden. Inhaltlich lassen sich alle Faktoren nachvollziehbar interpretieren. Das alles spricht für auf den ersten Blick zumindest hinreichend akzeptable Lösung des Faktorenmodells (vgl. Abschnitt 3.2.3). Es gibt jedoch auch mehrere Probleme:

- Fünf Faktoren haben jeweils weniger als drei Variablen, die auf diese Faktoren laden.
- Manche Variablen laden gleichzeitig auf mehrere Faktoren (drei der 24 Variablen).

Das entspricht nicht der „einfach Ladungsstruktur“, welche in der Literatur (Kieffer 1998; Abschnitt 3.2.3) besprochen wurde.

Bei dem Faktorenmodell mit Obliminrotation ist es nicht viel besser: Nur drei der acht Faktoren haben mindestens drei Variablen, welche mit >0.4 auf diese Faktoren laden. Die restlichen fünf Faktoren haben weniger als drei Variablen, welche mit >0.4 auf sie laden. Zwei der 24 Variablen laden auf mehrere Faktoren.

Insgesamt liefert die Faktorenanalyse mäßig aussagekräftigen Ergebnisse, welche z.T. mit erheblicher Unsicherheiten verbunden sind.

- Es herrscht keine Klarheit darüber, ob die Beobachtungen unabhängig sind (oder ob sie gar stark abhängig sind).
- Es mussten mehrere niedrigkorrelierte Variablen ausgeschlossen werden, damit die Korrelationsmatrix überhaupt zur Faktorenextraktion geeignet ist (Singularität).
- Analysiert man Faktorladungen >0.2 (anstatt >0.4) wird die Interpretation der Faktoren erheblich verkompliziert (vgl. Tabelle 11), sodass im Grunde Modellgenauigkeit zugunsten der Interpretierbarkeit aufgegeben werden musste.
- Das KMO-Kriterium und die vergleichsweise niedrigen MSA-Werte deuteten bereits vor der Faktorenanalyse an, dass die Korrelationsmatrix nicht besonders gut für eine Faktorenanalyse geeignet ist.
- Die beiden Faktorenmodelle mit acht Faktoren erklären lediglich 57% der Varianz des Datensatzes.
- Es gibt mehrere Variablen die auf mehrere Faktoren laden und die Mehrheit der Faktoren wird durch weniger als drei >0.4 ladende Variablen beschrieben.

Nichtsdestrotz konnte man durch die explorative Faktorenanalyse einige Erkenntnisse gewinnen: Es konnte gezeigt werden, dass sich mehrere Variablen sinnvoll zusammenfassen lassen können (vgl. z.B. die Faktoren f_1, f_2, f_3, f_4 *und* f_7), sodass Zusammenhänge zwischen den Variablen aufgedeckt werden konnten.

Eine Variablenreduktion durch eine Datentransformation mithilfe eines explorativen Faktorenmodells ist angesichts des erklärten Varianzanteils von 57% (vgl. Tabelle 10) aber fragwürdig, da ein erheblicher Teil der Streuung der Daten verloren gehen würde.

$Faktorenzahl$	$Varianzanteil_V$	$Kum.Varianzanteil_V$	$Varianzanteil_O$	$Kum.Varianzanteil_O$
1	0.14	0.14	0.12	0.12
2	0.11	0.25	0.10	0.22
3	0.08	0.33	0.07	0.29
4	0.06	0.39	0.08	0.37
5	0.06	0.44	0.07	0.44
6	0.05	0.49	0.05	0.49
7	0.05	0.54	0.05	0.54
8	0.03	0.57	0.04	0.57

Tabelle 10: Erklärte Varianzanteile der Faktorenmodelle. Das Unterzeichen $_V$ steht dabei für das Faktormodell mit Varimaxrotation, das Unterzeichen $_O$ steht für das Modell mit Oblimin-rotation. Quelle: Eigene Darstellung.

	f_1^v	f_2^v	f_3^v	f_4^v	f_5^v	f_6^v	f_7^v	f_8^v	u_v^2	h_v^2	f_1^o	f_2^o	f_3^o	f_4^o	f_5^o	f_6^o	f_7^o	f_8^o	u_o^2	h_o^2
num_self_hrefs							0.43		0.81	0.19						0.51			0.81	0.19
num_videos							0.21		0.93	0.07						0.29			0.93	0.07
num_imgs							0.35		0.84	0.16						0.35			0.84	0.16
num_hrefs		-0.28					0.67		0.42	0.58						0.71			0.42	0.58
n_non_stop_unique_tokens		0.87						0.27	0.12	0.88		0.90							0.12	0.88
n_unique_tokens		0.96							0.00	1.00		0.99							0.00	1.00
n_tokens_content		-0.65		-0.27			0.40		0.25	0.75		-0.50		-0.24					0.25	0.75
average_token_length		0.24					0.21	0.71	0.37	0.63								0.75	0.37	0.63
n_tokens_title									0.99	0.01									0.99	0.01
title_sentiment_polarity									0.96	0.04									0.96	0.04
title_subjectivity									0.96	0.04					0.21				0.96	0.04
LDA_00									0.95	0.05									0.95	0.05
global_sentiment_polarity	0.66		0.54	0.24	0.30				0.13	0.87	-0.42		0.52	0.22	0.32				0.13	0.87
rate_positive_words	0.90				0.24			0.31	0.03	0.97	-0.80				0.24			0.26	0.03	0.97
global_subjectivity			0.45	-0.25	0.22				0.65	0.35			0.37	-0.30					0.65	0.35
min_positive_polarity		0.39	0.27		-0.37				0.58	0.42		0.29	0.39		-0.39				0.58	0.42
global_rate_negative_words	-0.88				0.28				0.11	0.89	0.97				0.31				0.11	0.89
avg_negative_polarity	0.24			0.69		0.66			0.00	1.00				0.75			0.49		0.00	1.00
global_rate_positive_words	0.33				0.82				0.15	0.85					0.89				0.15	0.85
rate_negative_words	-0.96				-0.25				0.01	0.99	0.89				-0.26				0.01	0.99
min_negative_polarity	0.33	0.30		0.79					0.11	0.89				0.92					0.11	0.89
max_positive_polarity		-0.28	0.60		0.26				0.41	0.59		-0.24	0.52		0.21				0.41	0.59
max_negative_polarity		-0.22				0.81			0.28	0.72							0.83		0.28	0.72
avg_positive_polarity			0.91						0.16	0.84			0.93						0.16	0.84

Tabelle 11: Faktorladungen für varimax- und obliminrotierte Faktorenmodelle, wenn Faktorladungsbeträge >0.2 zur Untersuchung mit hinzugezogen werden. Quelle: Eigene Darstellung.

4.4 Hauptkomponentenanalyse

4.4.1 Vorbereitung: Anforderungen, Hauptkomponentenzahl, Rotation

Bei der Hauptkomponentenanalyse werden die selben Variablen wie bei der explorativen Faktorenanalyse verwendet, damit die Lösungen verglichen werden können. Auch bei der Hauptkomponentenanalyse besteht das Problem, dass einige Variablen erkennbare nichtlineare Zusammenhänge haben, was die Aussagekraft dieses Modells negativ beeinflussen kann.
Eine Parallelanalyse nach Horn (vgl. Abbildung 11) ergibt, dass man sieben Hauptkomponenten extrahieren kann. Die Ladungen der sieben Hauptkomponenten, jeweils die der Varimaxrotation (PCA^v) und Obliminrotation (PCA^o), sind in Tabelle 12 abgebildet.

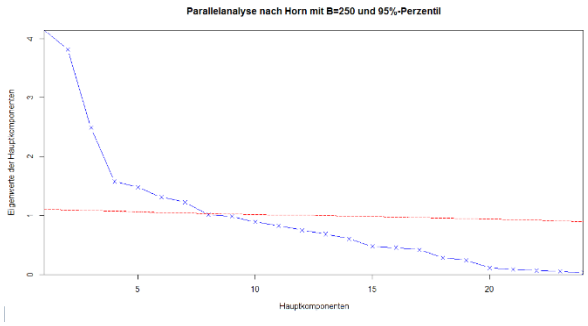

Abbildung 11: Ergebnis der Parallelanalyse nach Horn für die Hauptkomponentenanalyse. Es werden sieben Hauptkomponenten extrahiert. Die blaue Linie sind die Eigenwerte der Korrelationsmatrix der Daten, die rote Linie sind das 95%-Quantil der Eigenwerte der Korrelatonsmatrix der zufallsgenerierten Datensätze. Quelle: Eigene Darstellung.

4.4.2 Hauptkomponentenanalyse mit Varimaxrotation

Hauptkomponente *PC_1*: Diese Hauptkomponente ist dem Faktor f_1 der Variamaxrotation ähnlich (die Variablen, welche auf f_1 laden, laden auch auf PC_1 – mit ähnlichen Ladungen), und kann entsprechend interpretiert werden.

Hauptkomponente *PC_2*: Die Variablen *n_non_stop_unique_tokens*, *n_unique_tokens* und *average_token_length* laden positiv auf diese Hauptkomponente. Die Variable *n_token_content* lädt negativ auf diese Hauptkomponente. Diese Hauptkomponente lässt sich damit als Wortschatzkomplexitätsdichte interpretieren. Hat ein Artikel einen hohen Wert auf dieser Hauptkomponente, bedeutet das, dass der Artikel im Verhältnis zum Textumfang über einen hohen Wortschatz verfügt und dieser Wörter vergleichsweise lang sind.

Hauptkomponente *PC_3*: Diese Hauptkomponente ist dem Faktor f3 der Varimaxrotation ähnlich (die Variablen, welche auf f3 laden, laden auch auf PC3 – mit ähnlichen Ladungen), und kann entpsrechend interpretiert werden.

Hauptkomponente *PC₄*: Die Variablen *LDA_00* und *global_rate_positive_words* laden positiv auf diese Hauptkomponente und die Variable *min_positive_polarity* lädt negativ auf diese

	PCA_1^v	PCA_2^v	PCA_3^v	PCA_4^v	PCA_5^v	PCA_6^v	PCA_7^v	u_v^2	h_v^2	PCA_1^o	PCA_2^o	PCA_3^o	PCA_4^o	PCA_5^o	PCA_6^o	PCA_7^o	u_o^2	h_o^2
num_self_hrefs					0.71			0.47	0.53					0.75			0.47	0.53
num_videos								0.86	0.14								0.86	0.14
num_imgs					0.58			0.60	0.40					0.57			0.60	0.40
num_hrefs					0.72			0.33	0.67					0.71			0.33	0.67
n_non_stop_unique_tokens		0.88						0.16	0.84		0.90						0.16	0.84
n_unique_tokens		0.91						0.08	0.92		0.91						0.08	0.92
n_tokens_content		-0.58						0.25	0.75		-0.55						0.25	0.75
average_token_length		0.61						0.35	0.65		0.68						0.35	0.65
n_tokens_title								0.75	0.25								0.75	0.25
title_sentiment_polarity							0.64	0.55	0.45							0.66	0.55	0.45
title_subjectivity							0.80	0.35	0.65							0.82	0.35	0.65
LDA_00				0.55				0.67	0.33						0.60		0.67	0.33
global_sentiment_polarity	0.74		0.54					0.11	0.89	0.70		0.51					0.11	0.89
rate_positive_words	0.88							0.08	0.92	0.88							0.08	0.92
global_subjectivity			0.63					0.51	0.49			0.56					0.51	0.49
min_positive_polarity				-0.70				0.33	0.67						-0.64		0.33	0.67
global_rate_negative_words	-0.88							0.18	0.82	-0.89							0.18	0.82
avg_negative_polarity						0.90		0.07	0.93				0.92				0.07	0.93
global_rate_positive_words				0.56				0.34	0.66						0.54		0.34	0.66
rate_negative_words	-0.95							0.06	0.94	-0.96							0.06	0.94
min_negative_polarity						0.54		0.20	0.80				0.56				0.20	0.80
max_positive_polarity			0.73					0.31	0.69			0.65					0.31	0.69
max_negative_polarity						0.78		0.24	0.76				0.79				0.24	0.76
avg_positive_polarity			0.89					0.14	0.86			0.94					0.14	0.86

Tabelle 12: Ladungen (>0.5) der Variablen auf den Hauptkomponenten. Das hochgestellte *v* und *o* steht für Varimaxrotation bzw. Obliminrotation. Quelle: Eigene Darstellung.

Hauptkomponente. Hat ein Artikel ein hohen Wert auf dieser Hauptkomponente, bedeutet dass, dass der Artikel tendenziell eine hohe Nähe zum Thema LDA_00 hat und zahlreiche positive Wörter zu finden sind, welche jedoch einen niedrigen Minimumwert positiver Polarität haben. Diese Hauptkomponente ist schwer zu interpretieren (und benennen), da (a) nicht bekannt ist, was das Thema LDA_00 ausmacht (vgl. Abschnitt 2.2) und die Variable *min_positive_polarity* nur das Minimum der Polaritätswerte der positiv polaren Wörter angibt (was nicht zwangsläufig etwas über die Verteilung der positiven Polaritätswerte in dem Artikel aussagt).

Hauptkomponente PC_5: Die Variablen *num_hrefs*, *num_self_hrefs* und *num_imgs* laden positiv auf diese Hauptkomponente, welche sich als Anzahl der Bilder und Verlinkung des Artikels interpretieren lässt. Hat ein Artikel einen hohen Wert auf dieser Hauptkomponente, kann man erwarten, dass er im Schnitt viele Links und Bilder besitzt.

Hauptkomponente PC_6: Die Variablen *min_negative_polarity*, *avg_negative_polarity* und *max_negative_polarity* laden positiv auf diese Hauptkomponente. Sie kann man daher als Polaritätsstarke der negativ polaren Wörter interpretiert werden. Hat ein Artikel einen hohen Wert auf dieser Hauptkomponente, bedeutet das, dass die negativen Wörter im Artikel (unabhängig davon wie viele es sind) stark negativ polar sind.

Hauptkomponente PC_7: Die Variablen *title_sentiment_polarity* und *title_subjectivity* laden positiv auf diese Hauptkomponente, welche sich deshalb als Subjektivität und Polarität der Überschrift interpretieren lässt. Hat ein Artikel einen hohen Wert auf dieser Hauptkomponente, kann man erwarten, dass die Überschrift sehr polar und subjektiv ist.

4.4.3 Hauptkomponentenanalyse mit Obliminrotation

Wenn man nur Ladungen >0.5 betrachtet (vgl. Tabelle 12), sind die Hauptkomponenten PC_1, PC_2, PC_3, PC_5 und PC_7 nach einer Oblimin- und Varimaxrotation fast identisch (die Ladungsbeträge unterscheiden sich nur um wenige Prozent). Die Hauptkomponenten PC_4 und PCA_6 nach einer Varimaxrotation sind jeweils fast identisch mit den Hauptkomponenten PC_6 und PC_4 nach einer Obliminrotation. Insgesamt gibt es, wenn man Ladungen >0.5 betrachtet, zwischen den unterschiedlich rotiertn Modellen nur kleinere Unterschiede (wenige Prozent) bei den Faktorladungen. Die obliminrotierten Hauptkomponenten korrelieren (vgl. Tabelle 13) untereiander nur schwach (in der Regel <0.2).

	PCA_1	PCA_2	PCA_3	PCA_4	PCA_5	PCA_6	PCA_7
PCA_1	1.00	0.09	0.08	0.17	0.00	0.09	0.09
PCA_2	0.09	1.00	0.01	-0.01	-0.21	-0.18	-0.06
PCA_3	0.08	0.01	1.00	-0.04	0.13	0.19	0.12
PCA_4	0.17	-0.01	-0.04	1.00	-0.04	-0.06	-0.02
PCA_5	0.00	-0.21	0.13	-0.04	1.00	0.21	-0.00
PCA_6	0.09	-0.18	0.19	-0.06	0.21	1.00	0.14
PCA_7	0.09	-0.06	0.12	-0.02	-0.00	0.14	1.00

Tabelle 13: Korrelationen der obliminrotierten Hauptkomponenten. Quelle: Eigene Darstellung.

4.4.4 Vergleich und Bewertung der Hauptkomponentenmodelle

Beide Rotationsverfahren kommen zu ähnlichen Hauptkomponenten mit sehr ähnlichen Ladungsstrukturen und Ladungshöhen. Wie auch bei der explorativen Faktorenanalyse ist das zum Teil dem abschneiden der Faktorladungen geschuldet: Zieht man zur Interpretation Beispielsweise alle Faktorladungen mit einem Betrag >0.2 (anstatt >0.5) zur Analyse hinzu, unterscheiden sich die Lösungen der beiden Rotationsmethoden weitaus stärker (vgl. Tabelle 15).
Es wurde sich jedoch für eine Mindesthöhe der Ladungen >0.5 entscheiden, da geringere Mindesthöhen der Ladungen dazu führen würden, dass jeweils sehr viele Variablen gleichzeitig auf mehrere Hauptkomponenten laden.-Dadurch (und durch die vielen zusätzlichen kleinen Ladungen) würde die Interpretation der Hauptkomponenten erheblich verkompliziert werden.
Im Vergleich zur explorativen Faktorenanalyse ist die Ladungsstruktur „besser“ (vgl. Abschnitt 3.2,3): Bei den varimax- und obliminrotierten Hauptkomponentenmodellen (Ladungen >0.5) gibt es jeweils nur eine Variable, welche gleichzeitig auf mehrere Hauptkomponenten lädt. Fast alle Hauptkomponenten lassen sich inhaltlich sehr gut interpretieren.
Allerdings gibt es auch einige Probleme:

- Analysiert man Ladungen >0.2 (anstatt >0.5) wird die Interpretation der Hauptkomponenten erheblich verkompliziert, sodass im Grunde Modellgenauigkeit zugunsten der Interpretierbarkeit aufgegeben werden musste.
- Die beiden Hauptkomponentenmodelle mit jeweils sieben Hauptkomponenten erklären lediglich 67% der Varianz des Datensatzes.

$Hauptkomponentenzahl$	$Varianzanteil_V$	$Kum.Varianzanteil_V$	$Varianzanteil_O$	$Kum.Varianzanteil_O$
1	0.15	0.15	0.15	0.15
2	0.12	0.26	0.12	0.27
3	0.11	0.37	0.10	0.36
4	0.08	0.45	0.08	0.45
5	0.08	0.53	0.08	0.53
6	0.08	0.61	0.08	0.61
7	0.06	0.67	0.06	0.67

Tabelle 14: Erklärter Varianzanteil der Hauptkomponentenmodelle. Das untergestellte Zeichen *v* steht für das varimaxrotatierte Modell, *o* für das obliminrotierte. Quelle: Eigene Darstellung

	PCA_1^v	PCA_2^v	PCA_3^v	PCA_4^v	PCA_5^v	PCA_6^v	PCA_7^v	u_v^2	h_v^2	PCA_1^o	PCA_2^o	PCA_3^o	PCA_4^o	PCA_5^o	PCA_6^o	PCA_7^o	u_o^2	h_o^2
num_self_hrefs					0.71			0.47	0.53					0.75	-0.24		0.47	0.53
num_videos							0.31	0.86	0.14							0.30	0.86	0.14
num_imgs					0.58			0.60	0.40					0.57	-0.29		0.60	0.40
num_hrefs		-0.23		0.24	0.72			0.33	0.67					0.71			0.33	0.67
n_non_stop_unique_tokens		0.88			-0.23			0.16	0.84		0.90						0.16	0.84
n_unique_tokens		0.91		-0.21				0.08	0.92		0.91						0.08	0.92
n_tokens_content		-0.58	0.30	0.35	0.44			0.25	0.75		-0.55			0.37	0.25		0.25	0.75
average_token_length		0.61		0.33	0.34			0.35	0.65		0.68			0.43	0.35		0.35	0.65
n_tokens_title							0.44	0.75	0.25						-0.24	0.43	0.75	0.25
title_sentiment_polarity							0.64	0.55	0.45							0.66	0.55	0.45
title_subjectivity							0.80	0.35	0.65							0.82	0.35	0.65
LDA_00				0.55				0.67	0.33						0.60		0.67	0.33
global_sentiment_polarity	0.74		0.54					0.11	0.89	0.70		0.51					0.11	0.89
rate_positive_words	0.88			0.26				0.08	0.92	0.88							0.08	0.92
global_subjectivity			0.63			-0.28		0.51	0.49			0.56	-0.32				0.51	0.49
min_positive_polarity		0.31	0.21	-0.70				0.33	0.67		0.26	0.40			-0.64		0.33	0.67
global_rate_negative_words	-0.88							0.18	0.82	-0.89					0.23		0.18	0.82
avg_negative_polarity	0.29					0.90		0.07	0.93				0.92				0.07	0.93
global_rate_positive_words	0.40		0.37	0.56			0.22	0.34	0.66	0.38					0.54	0.23	0.34	0.66
rate_negative_words	-0.95							0.06	0.94	-0.96							0.06	0.94
min_negative_polarity	0.47	0.32	-0.29	-0.28		0.54		0.20	0.80	0.37	0.29		0.56		-0.26		0.20	0.80
max_positive_polarity		-0.22	0.73	0.28				0.31	0.69		-0.21	0.65			0.27		0.31	0.69
max_negative_polarity		-0.26				0.78		0.24	0.76	-0.31	-0.23		0.79		0.22		0.24	0.76
avg_positive_polarity			0.89	-0.24				0.14	0.86			0.94					0.14	0.86

Tabelle 15: Ladungen >0.2 der Hautkomponentenmodelle (Varimax- und Obliminrotation). Quelle: Eigene Darstellung.

4.5 Clusteranalyse

4.5.1 Durchführung der agglomerativen hierarchischen Clusteranalyse

Die hierarchische Clusterstruktur wird in Abbildung 10 mithilfe von zwei Dendrogrammen dargestellt. Die durchschnittlichen Silhouettenkoeffizienten wurden für diese Clusterstruktur in Abhängigkeit von der Clusterzahl *k=2, …,15* ermittelt. Auf Basis des Dendrogramms der z-standardisierten Daten wurde die Entscheidung getroffen, insgesamt fünf Cluster zu extrahieren.

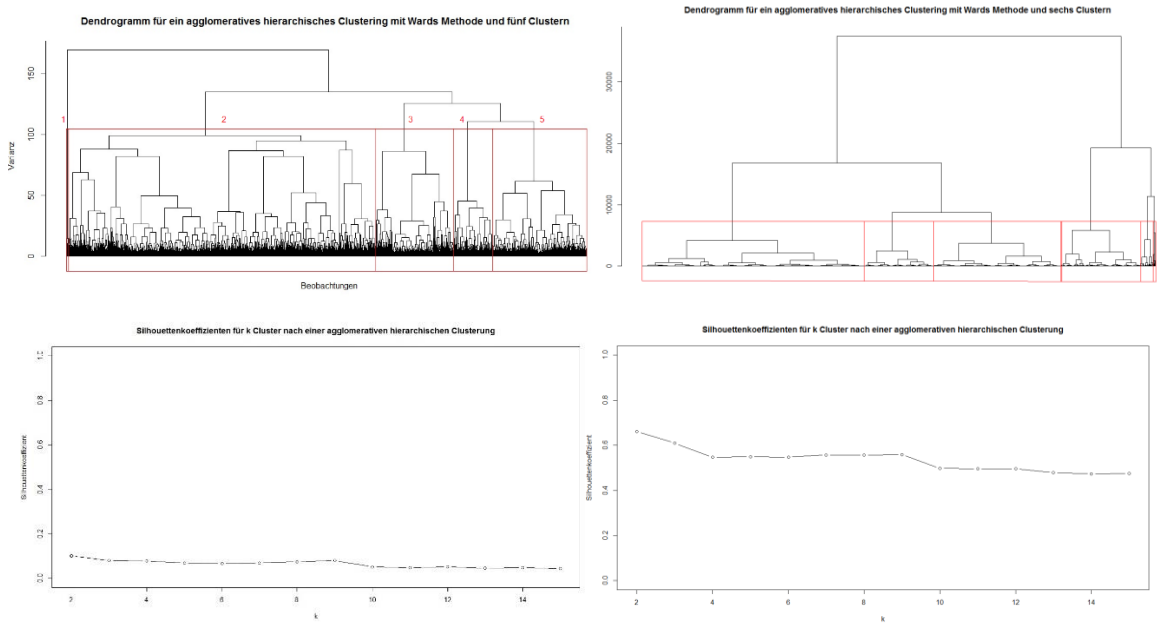

Abbildung 10: Dendrogramme (oben) und die durchschnittlichen Silhouettenkoeffizienten (unten) in Abhängigkeit von *k* Clustern (auf der linken Seite für die standardisierten Daten, auf der rechten Seite für die nichtstandardisierte Daten). Quelle: Eigene Darstellung.

In den nichtstandardisierten Daten existiert eine ausgeprägte Clusterstruktur, was man an den den durchschnittlichen Silhouettenkoeffizienten (sie liegen für verschiedene *k* um 0.6) und dem Dendrogramm (die Bäume haben in den ersten Verzweigungen einen starken Varianzunterschied) erkennen kann. Bei den standartisierten Daten ist das anders:
Die durchschnittlichen Silhouettenkoeffizienten für verschiedene *k* liegen um 0.1 - Cluster von-gleichverteilten Zufallsdatensätze erzeugen Werte um 0 (Rousseeuw 1987) - und im Dendrogramm kann man auch keinen erheblichen Varianzunterschied in den ersten Verzweigungen der Bäume erkennen.
Zur Analyse der Clusterstruktur wurden - getrennt nach Clustern - jeweils das erste Quartil (Q_1), der Median (Q_2), der Mittelwert (Mw) und das dritte Quartil (Q_3) für jeder Variable berechnet (vgl. Tabelle 16). Zur besseren Interpretation wurden diese Kennzahlen auf Basis der

nichtstandardisierten Daten berechnet (die Clusteranalyse wurde jedoch auf Basis der standardisierten Daten durchgeführt). Zusätzlich dazu wurden die Variablen aufsteigend nach den f-Werten der ANOVA (welche auf Basis der standardisierten Daten berechnet wurden) sortiert. Wenn man sich für jede Variable Clusterweise die Mediane und Mittelwerte anschaut und vergleicht, fällt auf, dass die Unterschiede zwischen den Clustern in der Regel gering sind (aber existieren). Es ist schwierig Variablen zu finden, die Cluster auf den ersten Blick erkennbar unterteilen oder auszeichnen. Das kann mehrere Gründe haben:

- Man verliert trotz der Sortierung nach den f-Werten den Überblick über die Menge der Zahlen. Immerhin sind 24 Variablen für fünf Cluster zu untersuchen, sodass ein Vergleich Anhand von Mittelwerten und Medianen die Betrachtung von 240 Kennzahlen erfordert.
- Die Clusterstruktur ist schwach ausgeprägt, sodass es keine augenscheinlichen Unterschiede gibt.

Ein Cluster (C_5) sticht jedoch hervor, da Quartile und Mittelwerte von vielen Variablen null sind. Dieses Cluster besteht aus 23 Beobachtungen, die – der Tabelle zu urteilen - offenbar fast niemals Wörter (Textinhalt) besitzen. Die URLs der Artikel wurden herausgesucht (vgl. Anhang) – und es stellte sich heraus, dass es sich um normale Artikel mit Text und Bilder handelt. Vermutlich gab es bei der Extraktion dieser Artikel Probleme, sodass die Kennzahlen nicht richtig berechnet werden konnten. Diese Artikel wurden aus dem Datensatz entfernt, und es wurde eine erneute agglomerative hierarchische Clusteranalyse durchgeführt, welche jedoch keine erhebliche Verbesserung des durchschnittlichen Silhouettenkoeffizienten nach sich zog (vgl. Abbildung 11). Eine inhaltlich sinnvolle Interpretation der Cluster war damit nicht möglich (vgl. Abschnitt 4.5.4).

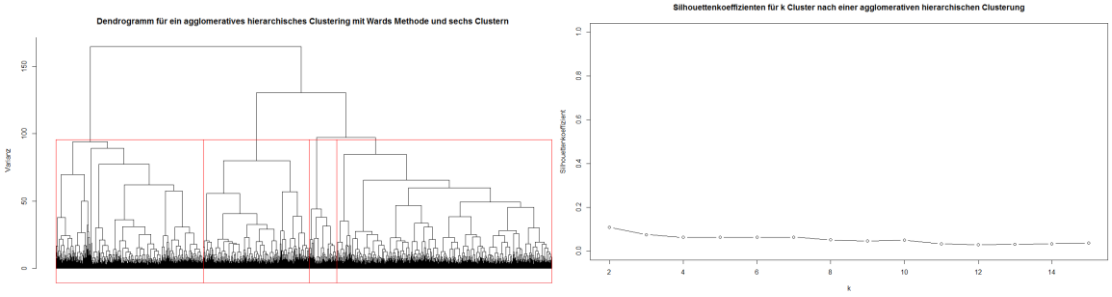

Abbildung 11: Die hierarchische Clusterstruktur und die durchschnittlichen Silhouettenkoeffizienten für den standardisierten Datensatz, der von den fehlerhaft erfassten Artikeln bereinigt wurde. Quelle: Eigene Darstellung.

	f	Q_1^1	Q_2^1	Mw^1	Q_3^1	Q_1^2	Q_2^2	Mw^2	Q_3^2	Q_1^3	Q_2^3	Mw^3	Q_3^3	Q_1^4	Q_2^4	Mw^4	Q_3^4	Q_1^5	Q_2^5	Mw^5	Q_3^5
n_tokens_title	5.58	9.00	10.00	10.26	12.00	9.00	10.00	10.37	12.00	9.00	10.00	10.20	12.00	9.00	11.00	10.78	12.75	9.50	11.00	10.74	12.50
num_self_hrefs	12.63	1.00	2.00	2.72	4.00	1.00	2.00	2.55	4.00	1.00	2.00	3.07	4.00	2.00	3.00	2.68	4.00	0.00	0.00	0.00	0.00
num_videos	37.23	0.00	0.00	0.28	0.00	0.00	0.00	0.28	0.00	0.00	0.00	1.39	0.00	0.00	0.00	0.25	0.00	1.00	1.00	0.87	1.00
title_subjectivity	75.37	0.00	0.00	0.20	0.40	0.00	0.00	0.19	0.40	0.00	0.00	0.34	0.50	0.00	0.27	0.24	0.38	0.00	0.07	0.33	0.59
LDA_00	116.31	0.47	0.65	0.63	0.80	0.42	0.64	0.60	0.80	0.63	0.80	0.73	0.87	0.41	0.56	0.57	0.74	0.41	0.47	0.51	0.59
num_imgs	147.23	1.00	1.00	1.14	1.00	1.00	1.00	1.14	1.00	1.00	1.00	3.28	2.00	0.00	1.00	0.74	1.00	0.00	0.00	0.09	0.00
max_negative_polarity	159.24	-0.15	-0.10	-0.13	-0.07	-0.12	-0.10	-0.12	-0.07	-0.12	-0.08	-0.09	-0.05	0.00	0.00	-0.03	0.00	0.00	0.00	0.00	0.00
title_sentiment_polarity	215.02	0.00	0.00	0.03	0.07	0.00	0.00	-0.01	0.00	0.00	0.10	0.20	0.40	0.00	0.00	0.05	0.10	0.00	0.00	0.10	0.20
avg_positive_polarity	247.88	0.30	0.34	0.35	0.39	0.25	0.31	0.31	0.37	0.34	0.38	0.38	0.41	0.31	0.38	0.37	0.43	0.00	0.00	0.00	0.00
min_positive_polarity	278.44	0.03	0.10	0.09	0.10	0.06	0.10	0.14	0.14	0.03	0.03	0.06	0.10	0.06	0.10	0.14	0.17	0.00	0.00	0.00	0.00
global_subjectivity	289.13	0.38	0.43	0.43	0.48	0.34	0.39	0.40	0.45	0.42	0.46	0.46	0.51	0.36	0.42	0.42	0.48	0.00	0.00	0.00	0.00
min_negative_polarity	376.23	-0.60	-0.40	-0.45	-0.25	-0.70	-0.50	-0.50	-0.30	-0.80	-0.60	-0.59	-0.40	0.00	0.00	-0.04	0.00	0.00	0.00	0.00	0.00
num_hrefs	404.68	4.00	6.00	7.18	9.00	3.00	5.00	5.82	7.00	7.00	12.00	14.79	19.00	4.00	5.00	5.21	6.00	0.00	0.00	0.00	0.00
avg_negative_polarity	414.63	-0.30	-0.23	-0.25	-0.18	-0.32	-0.25	-0.27	-0.19	-0.30	-0.25	-0.26	-0.20	0.00	0.00	-0.03	0.00	0.00	0.00	0.00	0.00
global_rate_negative_words	500.69	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.01	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
global_rate_positive_words	562.24	0.03	0.04	0.04	0.05	0.02	0.02	0.03	0.03	0.04	0.05	0.05	0.06	0.03	0.05	0.05	0.06	0.00	0.00	0.00	0.00
n_tokens_content	656.31	240.00	344.00	423.55	546.00	192.00	257.50	298.76	357.00	508.00	803.00	871.85	1061.00	118.00	160.00	176.28	216.75	0.00	0.00	0.00	0.00
max_positive_polarity	813.13	0.60	0.75	0.74	1.00	0.50	0.50	0.54	0.60	0.80	1.00	0.91	1.00	0.50	0.70	0.68	0.80	0.00	0.00	0.00	0.00
n_unique_tokens	946.46	0.51	0.57	0.57	0.62	0.55	0.60	0.60	0.65	0.42	0.47	0.48	0.53	0.62	0.67	0.67	0.73	0.00	0.00	0.00	0.00
global_sentiment_polarity	977.22	0.09	0.13	0.13	0.17	-0.01	0.02	0.02	0.05	0.13	0.17	0.18	0.21	0.16	0.22	0.22	0.27	0.00	0.00	0.00	0.00
n_non_stop_unique_tokens	1034.26	0.67	0.72	0.72	0.77	0.69	0.74	0.74	0.79	0.61	0.65	0.65	0.70	0.75	0.80	0.80	0.87	0.00	0.00	0.00	0.00
rate_negative_words	1325.53	0.18	0.25	0.26	0.33	0.40	0.46	0.48	0.54	0.15	0.22	0.22	0.28	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
rate_positive_words	1606.69	0.67	0.75	0.74	0.82	0.46	0.53	0.52	0.60	0.72	0.78	0.78	0.85	1.00	1.00	0.96	1.00	0.00	0.00	0.00	0.00
average_token_length	1781.86	4.53	4.69	4.70	4.85	4.53	4.70	4.70	4.86	4.52	4.68	4.72	4.87	4.54	4.71	4.71	4.86	0.00	0.00	0.00	0.00

Tabelle 16: Verteilung der Variablenausprägungen nach einer agglomerativen hierarchischen Clusteranalyse (nach Clustern getrennt und nach f-Werten sortiert). Quelle: Eigene Darstellung

4.5.2 Durchführung der Clusteranalyse mit K-means

Um den Parameter k zu bestimmen, wurden Silhouettenkoeffizienten und erklärte Varianzanteile für *k=2, …,15* Cluster berechnet (jeweils für die standardisierten und nichtstandardisierten Daten) und in Abbildung 12 dargestellt:

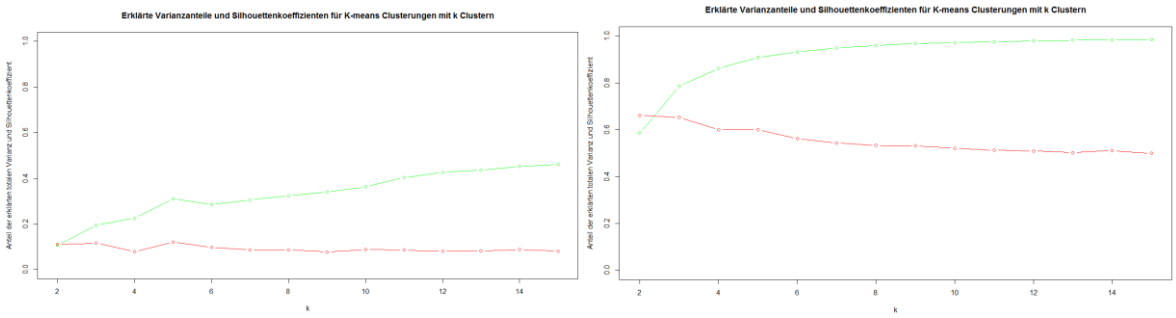

Abbildung 12: Durch die Clusterungen erklärte Varianzanteile (grün) und Silhouettenkoeffizienten (rot) für K-means Clusterungen mit *k* Clustern. Im Linken Bild sind die Daten standartisiert, im rechten nicht.

Wie schon beim hierarchischen Clustering kann man erkennen: Nach den technischen Kriterien (hier: Erklärter Varianzanteil und durchschnittliche Silhouettenkoeffizienten) gibt es in den nichtstandardisierten Daten eine ausgeprägte Clusterstruktur (die durchschnittlichen Silhouettenkoeffizienten liegen um 0.6 und der erklärte Varianzanteil geht fast gegen 100%), während die Clusterstruktur in den standartisierten Daten, wie schon beim agglomerativen hierarchischen Clustering, nur schwach ausgeprägt ist.

Zur Analyse der Clusterstruktur wurden auch für die K-Means Clusterlösung jeweils das erste Quartil (Q_1), der Median (Q_2), der Mittelwert (Mw) und das dritte Quartil (Q_3) für jedes Cluster auf jeder (nichtstandardisierten) Variable berechnet (vgl. Tabelle 17). Die Variablen wurden wieder aufsteigend nach den f-Werten der ANOVA (berechnet auf Basis der standardisierten Daten) sortiert.

Hier wiederholt sich, was bereits bei der agglomerativen hierarchischen Clusteranalyse zum Vorschein kam: Wenn man sich für jede Variable clusterweise die Mediane und Mittelwerte anschaut, fällt auf, dass die Unterschiede zwischen den Clustern nicht besonders groß sind (aber existieren). Auch hier ist es schwierig Variablen zu finden, die Cluster auf den ersten Blick erkennbar unterteilen oder auszeichnen. Eine inhaltlich sinnvolle Interpretation der Cluster war damit nicht möglich (vgl. Abschnitt 4.5.4).

	f	Q_1^1	Q_2^1	Mw^1	Q_3^1	Q_1^2	Q_2^2	Mw^2	Q_3^2	Q_1^3	Q_2^3	Mw^3	Q_3^3	Q_1^4	Q_2^4	Mw^4	Q_3^4	Q_1^5	Q_2^5	Mw^5	Q_3^5
n_tokens_title	15.05	9.00	10.00	10.34	12.00	9.00	10.00	10.06	11.00	9.00	10.00	10.19	12.00	9.00	11.00	10.70	12.00	9.00	10.00	10.36	12.00
average_token_length	34.68	4.52	4.69	4.63	4.84	4.50	4.64	4.65	4.79	4.57	4.73	4.75	4.91	4.54	4.70	4.76	4.93	4.50	4.67	4.68	4.78
num_self_hrefs	61.46	1.00	2.00	2.58	4.00	1.00	3.00	3.59	5.00	1.00	2.00	2.58	4.00	1.00	2.00	2.20	3.00	2.75	3.00	5.52	5.00
LDA_00	68.66	0.44	0.63	0.60	0.80	0.55	0.77	0.70	0.87	0.51	0.68	0.64	0.84	0.58	0.78	0.71	0.87	0.53	0.64	0.64	0.81
max_negative_polarity	79.66	-0.12	-0.10	-0.12	-0.07	-0.10	-0.07	-0.08	-0.05	-0.17	-0.12	-0.13	-0.05	-0.12	-0.10	-0.12	-0.05	-0.12	-0.10	-0.10	-0.05
avg_positive_polarity	110.90	0.27	0.33	0.33	0.38	0.34	0.37	0.37	0.41	0.30	0.34	0.35	0.40	0.34	0.38	0.38	0.42	0.38	0.40	0.40	0.42
num_imgs	132.70	1.00	1.00	1.23	1.00	1.00	1.00	3.44	3.00	1.00	1.00	1.09	1.00	1.00	1.00	1.49	1.00	0.00	1.00	1.59	1.00
global_subjectivity	139.96	0.35	0.41	0.41	0.46	0.42	0.46	0.46	0.50	0.37	0.43	0.43	0.48	0.41	0.46	0.46	0.51	0.45	0.50	0.50	0.55
min_positive_polarity	179.11	0.05	0.10	0.11	0.14	0.03	0.03	0.06	0.10	0.03	0.10	0.10	0.14	0.03	0.05	0.07	0.10	0.03	0.05	0.06	0.10
avg_negative_polarity	264.62	-0.32	-0.25	-0.27	-0.19	-0.31	-0.26	-0.27	-0.22	-0.23	-0.17	-0.18	-0.12	-0.31	-0.25	-0.26	-0.19	-0.38	-0.35	-0.35	-0.33
num_hrefs	517.14	4.00	6.00	6.29	8.00	8.00	12.00	15.41	19.00	4.00	5.00	6.16	8.00	5.00	9.00	9.88	13.00	13.25	25.00	28.11	30.00
n_non_stop_unique_tokens	523.70	0.69	0.73	0.73	0.78	0.60	0.63	0.63	0.67	0.70	0.75	0.75	0.80	0.65	0.69	0.70	0.74	0.56	0.61	0.61	0.68
title_subjectivity	526.67	0.00	0.00	0.21	0.40	0.00	0.00	0.17	0.30	0.00	0.00	0.17	0.38	0.45	0.50	0.59	0.75	0.18	0.50	0.46	0.67
global_rate_positive_words	567.69	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.05	0.03	0.04	0.05	0.06	0.04	0.05	0.05	0.06	0.05	0.06	0.05	0.06
max_positive_polarity	702.09	0.50	0.60	0.62	0.80	0.80	1.00	0.92	1.00	0.50	0.70	0.71	0.89	0.80	1.00	0.87	1.00	1.00	1.00	0.96	1.00
global_rate_negative_words	755.18	0.01	0.02	0.02	0.03	0.01	0.02	0.02	0.02	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02
title_sentiment_polarity	843.21	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.07	0.25	0.42	0.40	0.50	0.00	0.10	0.17	0.50
min_negative_polarity	961.89	-0.70	-0.50	-0.51	-0.30	-0.90	-0.70	-0.69	-0.50	-0.31	-0.20	-0.24	-0.12	-0.67	-0.50	-0.51	-0.31	-0.82	-0.76	-0.73	-0.60
n_unique_tokens	1005.00	0.54	0.58	0.58	0.63	0.41	0.45	0.45	0.48	0.56	0.60	0.61	0.66	0.48	0.53	0.54	0.58	0.44	0.51	0.50	0.54
global_sentiment_polarity	1182.33	0.02	0.06	0.05	0.09	0.11	0.14	0.15	0.18	0.13	0.17	0.18	0.22	0.14	0.18	0.18	0.22	0.12	0.15	0.15	0.20
n_tokens_content	1376.23	216.00	294.00	334.70	422.75	706.00	925.00	1012.50	1151.00	183.25	260.00	291.62	367.00	326.00	517.00	553.99	741.00	762.50	878.50	1159.75	1160.50
rate_negative_words	1620.14	0.33	0.38	0.40	0.47	0.20	0.26	0.26	0.31	0.09	0.15	0.15	0.21	0.14	0.20	0.20	0.26	0.20	0.23	0.25	0.28
rate_positive_words	1681.68	0.52	0.61	0.59	0.67	0.69	0.74	0.74	0.80	0.79	0.85	0.85	0.91	0.74	0.80	0.80	0.86	0.72	0.77	0.75	0.80
num_videos	2010.03	0.00	0.00	0.33	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.76	0.00	20.00	21.00	31.00	23.75

Tabelle 17: Verteilung der Variablenausprägungen nach einer K-Means Clusteranalyse (nach Clustern getrennt und nach f-Werten sortiert). Quelle: Eigene Darstellung

4.5.3 Durchführung der Clusteranalyse mit OPTICS

Für eine Clusteranalyse mit OPTICS mussten zuerst die Eingabeparameter ε und *MinPts* bestimmt werden. Dazu wurden diese Werte zuerst experimentell festgesetzt: ε wurde auf 100 und *MinPts* wurde jeweils einmal auf 24,48 und 72 gesetzt. Anschließend wurde eine Clusteranalyse mit OPTICS mit diesen drei Kombinationen auf den standardisierten Daten durchgeführt (vgl. Abbildung 13).

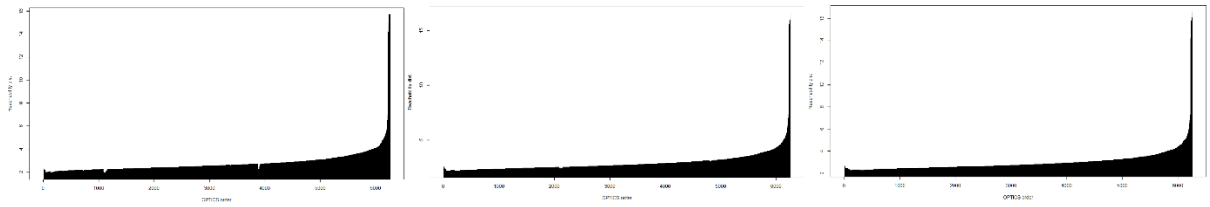

Abbildung 13: Erreichbarkeitsplot für OPTICS auf die standardisierten Daten mit ε=100 und *MinPts*-Werten 24 (links), 48 (mitte), 72 (rechts). Quelle: Eigene Darstellung.

In Abbildung 13 kann man erkennen, dass man ε zur besseren Übersicht auf die Erreichbarkeitsdistanzen auf einen kleineren Wert setzen kann. Folglich wurden drei weitere Durchgänge mit OPTICS durchgefährt, jeweils mit ε=5 und *MinPts*-Werten von 24,48 und 72 (vgl. Abbildung 14).

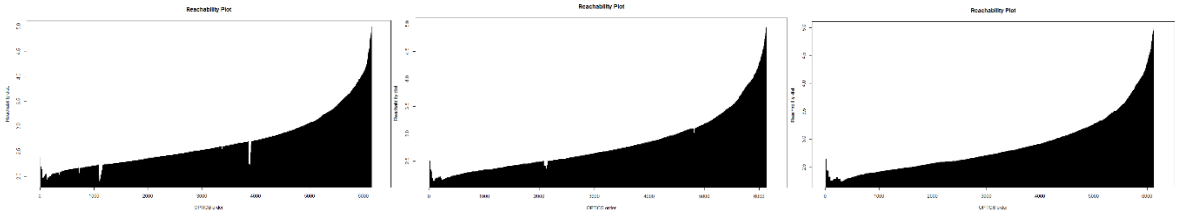

Abbildung 14: Erreichbarkeitsplot für OPTICS auf den standardisierden Daten mit ε=5 und *MinPts*-Werten 24 (links), 48 (mitte), 72 (rechts). Quelle: Eigene Darstellung.

Der zweite Durchlauf hat gezeigt, dass auf den standartisierten Daten kaum Cluster zu erkennen sind - insbesondere wenn man *MinPts* auf einen höheren Wert setzt. Zum Vergleich drei Durchläufe von OPTICS auf den nichtstandardisierten Daten (vgl. Abbildung 15):

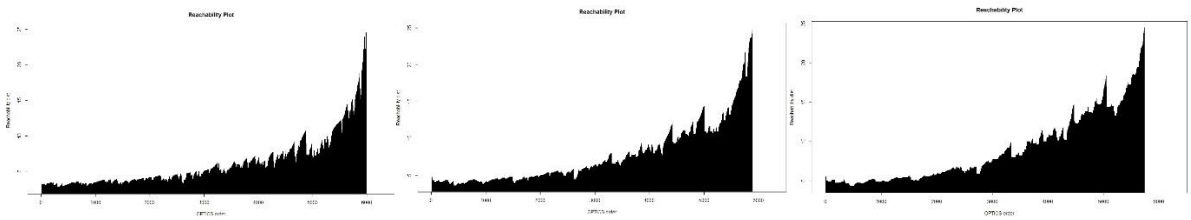

Abbildung 15: Die Erreichbarkeitsplots für drei Durchläufe mit Optics (ε=25 und *MinPts* 24 (links), 48 (mitte), 72 (rechts). Quelle: Eigene Darstellung.

An dem zackigen Verlauf der Erreichbarkeitsdistanzen (vgl. Abschnitt 3.4.3) in Abbildung 15 kann man erkennen, dass in den nichtstandardisierten Daten eine Clusterstruktur exisitert. In den standardisierten Daten konnte jedoch mithilfe des Erreichbarkeitsplots keine nennenswerte Clusterstruktur entdeckt werden, weswegen keine Cluster extrahiert wurden.

4.5.4 Bewertung der Clusteranalyse

Es wurden frei zufallsgenerierte Datensätze (ZGD) aus jeweils 6258 Beobachtungen und 24 unabhängigen, gleichverteilten (Wertebreich: [-1,1]) Zufallsvariablen erzeugt, um die Clusterlösungen der standardisierten Daten gegen die Clusterlösungen der ZGD antreten zu lassen. Bewertet wurde Anhand des erklärten Varianzanteils (K-Means), des durchschnittlichen Silhouettenkoeffizienten (agglomeratives hierarchisches Clusterung und K-Means und des Erreichbarkeitsplots (OPTICS).

Die Ergebnisse sind in den Abbildungen 16, 17 und 18 zu sehen. Bei der K-means-Clusterung hat der standardisierte Datensatz höhere erklärte Varianzen und durchschnittlichen Silhouettenkoeffizienten als die Clusterlösungen der drei ZGD. Nach einer Clusterung mit OPTICS zeigte, sich, dass die geordneten Erreichbarkeitsdistanzen des standartisierten Datensatzes und die der drei ZGD einen sehr ähnlichen Verlauf haben. Bei der agglomerativen hierarchischen Clusterung sind die durchschnittlichen Silhouettenkoeffizienten des standartisierten Datensatzes nur geringfügig größer als die der drei geclusterten ZGD.

Die Clusterstruktur in den standardisierten Daten ist äußerst schwach (vgl. Abschnitte 4.5.1 bis 4.5.3) – und, deskriptiv betrachtet, nur geringfügig besser als die der drei ZGD. Die Aussagekraft und der Nutzen einer Clusteranalyse kann deshalb als fragwürdig angesehen werden.

Die nichtstandardisierten Daten haben dagegen eine ausgeprägte Clusterstruktur, was aber vermutlich auf die sehr unterschiedlichen Wertebereiche der einzelnen Variablen zurückzuführen ist (vgl. Abschnitt 4.1). Die Aussagekraft einer Clusteranalyse auf den nichtstandardisierten Daten kann deshalb ebenfalls als fragwürdig angesehen werden.

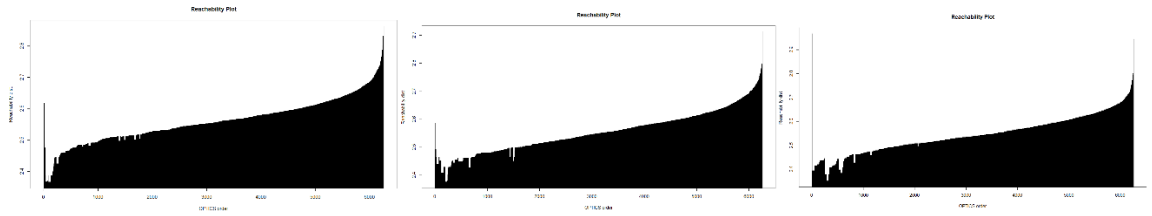

Abbildung 16: Der Erreichbarkeitsplot nach einer Analyse OPTICS (ε *=5, MinPts=48*) auf drei jeweils neu generierten ZGD. Quelle: Eigene Darstellung

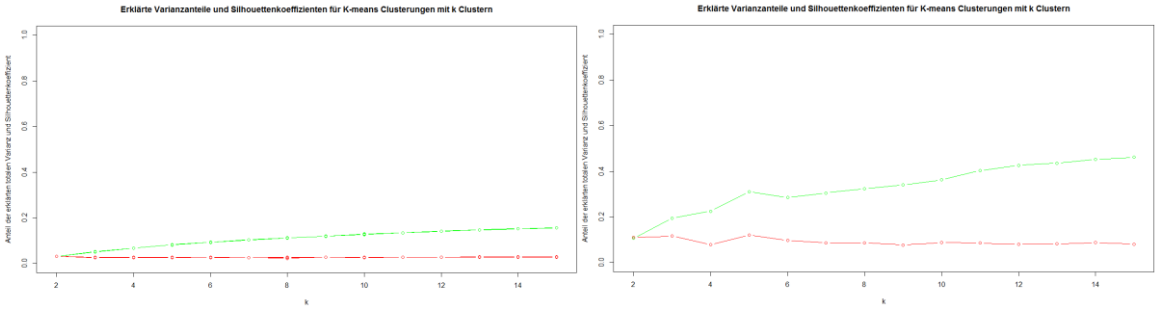

Abbildung 17: Vergleich ZGD mit den standardisierten Daten. Links wurden drei Mal für *k=2, ... 15* K-means Clusterungen durchgeführt und durchschnittliche Silhouettenkoeffilzentn und erkklärte Varianzanteile berechnet. Rechts die selben Kennzahlen für den standardisierten Datensatz. Quelle: Eigene Abbildung.

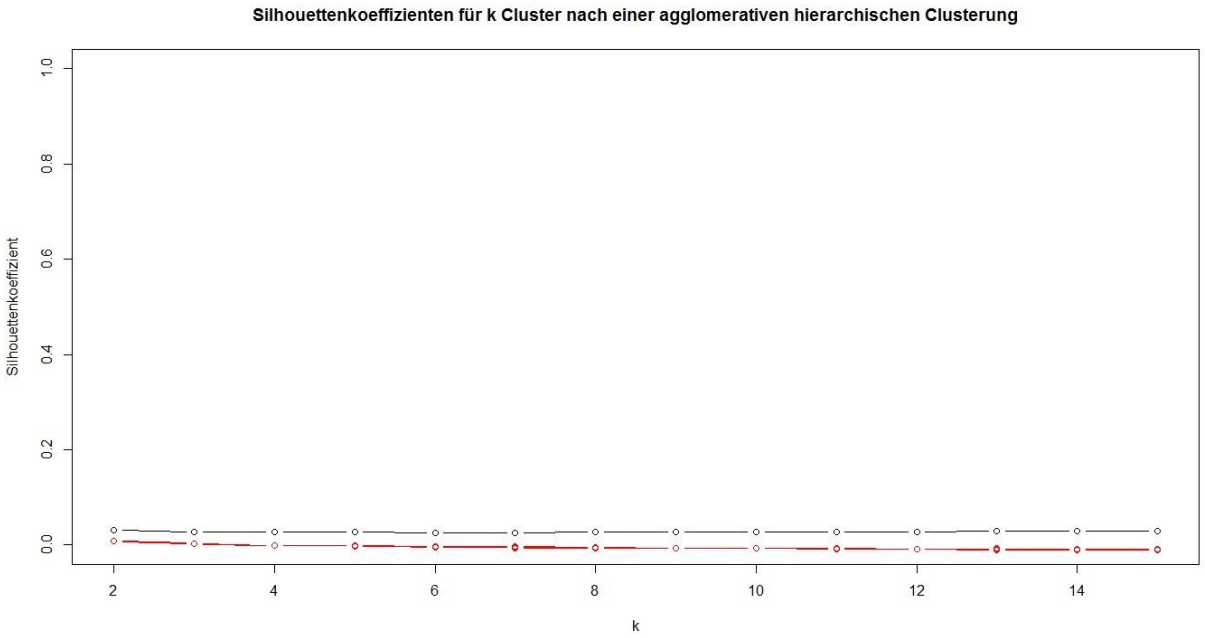

Abbildung 18: Die durchschnittlichen Silhouettenkoeffizienten nach agglomerativer hierarchischer Clusterung mit k Clustern (rot=ZGD,schwarz=standartisierter Datensatz). Quelle: Eigene Darstellung.

5 Fazit

Ziel dieser Arbeit war die Beantwortung dieser Fragen: Kann man Variablen, welche die Texte der Wirtschaftsnachrichten von Mashable beschreiben, effizient zusammenfassen (Variablenreduktion bzw. Dimensionsreduktion)? Stehen hinter manchen Variablen nichtbeobachtbare (latente) oder nicht erfasste Faktoren? Wie ist die Beziehung zwischen den erfassten Variablen? Gibt es Gruppierungen in den Texten Wirtschaftsnachrichten (z.B. Gruppen mit ähnlichem Wortschatz und Sprachstil)? Wenn ja – was zeichnet diese aus?

Antworten auf diese Fragen zu finden war schwerer, als es auf den ersten Blick vielleicht aussehen mag, da es erhebliche Probleme in der Datenqualität gibt:

- Es ist nicht klar, ob die Beobachtungen (die Artikel) überhaupt unabhängig sind (und falls sie nicht unabhängig sind - in welchem Ausmaß sind sie abhängig?).
- Es wurde gezeigt, dass sowohl Variablen als auch einige Beobachtungen von Fernandes et al. (2015a und b) falsch kalkuliert wurden.
- Bei mehreren Variablen haben Fernandes et al. (2015b) darüber hinaus nicht exakt beschrieben, wie sie kalkuliert wurden (z.B. die Anzahl der Tokens).
- Einige Variablen (*LDA00 – LDA04*) wurden nicht umfassend genug beschrieben, sodass mit diesen nicht angemessen gearbeitet werden konnte.

Trotz dieser Hindernisse konnten einige Erkenntnisse gewonnen werden:

- Die Korrelationsstruktur der untersuchten Daten erlaubt es grundsätzlich, dass man einige Variablen durch neue Faktoren (bzw. Hauptkomponenten) sinnvoll zusammenfassen kann (vgl. Abschnitt 4.3 und 4.4).
- Falls die Artikel inhaltlich unabhängig sein sollten, ist die explorative Faktorenanalyse und die Hauptkomponentenanalyse geeignet, um erste Beziehungen zwischen den Variablen herauszuarbeiten und potentiell zusammenfassbare Variablen *aufzudecken*.
- Da die Ergebnisse der explorativen Faktoren- und Hauptkomponentenanalyse jedoch unter schwierig interpretierbaren Faktorladungsstrukturen und geringen erklärten Varianzanteilen leiden, ist eine *Variablenreduktion* durch diese Modelle nicht angebracht.
- Die Clusterstruktur in den Wirtschaftsnachrichten ist schwach ausgeprägt, und der inhaltliche Nutzen einer Clusteranalyse erscheint – zumindest mit den untersuchten Daten – fragwürdig.
- Es konnte aufgedeckt werden, dass Fernandes et al. (2015) bei der Kalkulation einiger Variablen und Beobachtungen Fehler unterlaufen sind.

Die Analyse dieser Arbeit eröffnet eine Reihe von Fragestellungen, welche in zukünftigen Studien untersucht werden können:

Eine Frage wäre, ob die Clusterstruktur der standardisierten Daten der Wirtschaftsnachrichten von Mashable (auf statistisch signifikantem Ausmaß) ausgeprägter sind als die von gleichverteilt zufallsgenerierten Daten. Eine weitere Frage wäre, ob man mithilfe einer konfirmatorischen Faktoranalyse versuchen kann, einige der entdeckten Faktoren einzeln zu schätzen. Der Vorteil davon wäre, dass man gezielt versucht nur einige Variablen transformieren und dadurch möglicherweise eine vertretbare *Variablenreduktion* erreichen kann.

6 Anhang

6.1 Weitere Variablenklassen

Artikelkategorien. Die Variablen der Klasse „Artikelkategorien“ geben an, in welcher der sechs Artikelkategorien (Lifestyle, Unterhaltung, Wirtschaftsnachrichten, Social Media, Technik und weltweite Nachrichten) der Artikel veröffentlicht wurde. Da in dieser Arbeit ausschließlich die Wirtschaftsnachrichten untersucht werden, sind diese Variablen – sobald die Filterung nach den Wirtschaftsnachrichten stattgefunden hat - für die Analyse nicht relevant und werden im praktischen Teil aus dem Datensatz entfernt.

Variablenkürzel	Variablenbeschreibung	Skalenniveau	Wertebereich
data_channel_is_lifestyle	Der Artikel gehört gehört in die Kategorie Lifestyle	Binär	{0,1}
data_channel_is_entertainment	Der Artikel gehört gehört in die Kategorie Unterhaltung	Binär	{0,1}
data_channel_is_bus	Der Artikel gehört gehört in die Kategorie Wirtschaft	Binär	
data_channel_is_socmed	Der Artikel gehört gehört in die Kategorie Social Media	Binär	{0,1}
data_channel_is_tech	Der Artikel gehört gehört in die Kategorie Technisches	Binär	{0,1}
data_channel_is_world	Der Artikel gehört gehört in die Kategorie weltweite Nachrichten	Binär	{0,1}

Tabelle 18: Variablenbeschreibung, Skalenniveau und Wertebereich für die Variablen der Klasse „Artikelkategorie“.

Shares. Variablen der Klasse „Shares“ geben an, wie oft ein Artikel in einem sozialen Netzwerk von einem Leser verbreitet wird. Ein Leser kann einen Share auslösen, indem er mit Hilfe eines Internetbrowsers in einem sozialen Netzwerk angemeldet ist und auf der Website auf einen entsprechenden Knopf drückt. Fernandes, Vinagre und Cortez (2015) haben ausschließlich Shares der sozialen Netzwerke Facebook (www.facebook.com), Twitter (www.twitter.com), Pinterest (www.pinterest.com), LinkedIn (www.linkedin.com), StumbleUpon (www.stumbleupon.com) und Google Plus (www.plus.google.com) ermittelt. Shares sind durch Leser beeinflusste Werte und beeinflussen nicht den Text als solchen. Damit sind alle Kennzahlen der Klasse „Shares“ für den praktischen Teil dieser Arbeit nicht relevant und werden vor der Datenanalyse aus dem Datensatz entfernt.

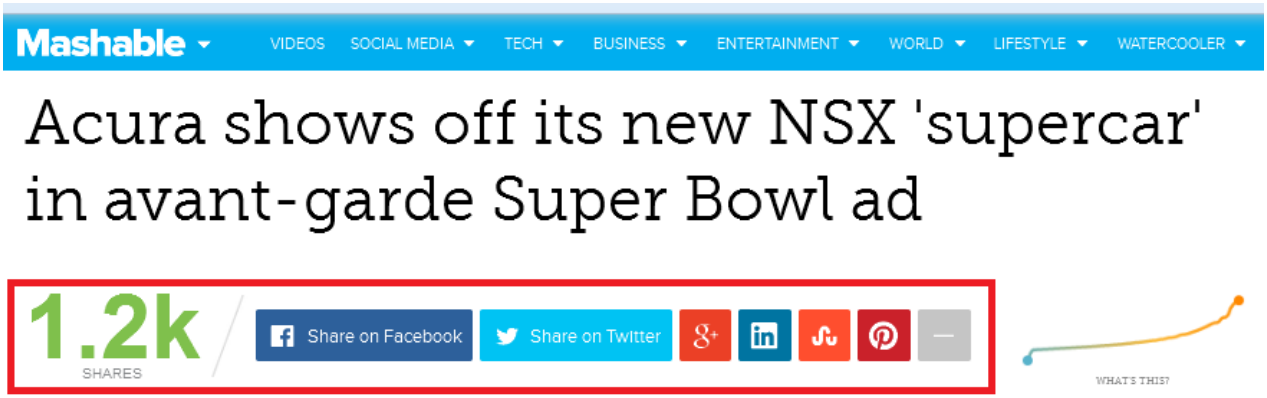

Abbildung: Das typische Mashable-Artikeldesign. Direkt unter der Unterschrift (im roten Kasten) kann man den Artikel in den sozialen Netzwerken teilen. Der rote Kasten wurde zur besseren Erkennung manuell eingefügt. Quelle: http://mashable.com/2016/01/29/acura-nsx-super-bowl-ad/#FCmGPRcL8uqo. Zuletzt aufgerufen am 30.01.2016.

Variablenkürzel	Variablenbeschreibung	Skalenniveau	Wertebereich
kw_min_min	Wie oft wurde das schlechteste Keyword mindestens geteilt?	Metrisch	[0,a]
kw_max_min	Wie oft wurde das schlechteste im Schnitt geteilt?	Metrisch	[0,a]
kw_avg_min	Wie oft wurde das schlechteste Keyword maximal geteilt?	Metrisch	[0,b]
kw_min_max	Wie oft wurde das beste mindestens geteilt?	Metrisch	[0,a]
kw_max_max	Wie oft wurde das beste im Schnitt geteilt?	Metrisch	[0,a]
kw_avg_max	Wie oft wurde das beste Keyword maximal geteilt?	Metrisch	[0,b]
kw_min_avg	Wie oft wurde das mittelmäßige mindestens geteilt?	Metrisch	[0,a]
kw_max_avg	Wie oft wurde das mittelmäßige im Schnitt geteilt?	Metrisch	[0,a]
kw_avg_avg	Wie oft wurde das mittelmäßige Keyword maximal geteilt?	metrisch	[0,b]
self_reference _min_shares	Anzahl Shares des verlinkten Artikels von Mashable mit den wenigsten Shares	metrisch	[0,a]
self_reference_ max_shares	Anzahl Shares des verlinkten Artikels von Mashable mit den meisten Shares	metrisch	[0,a]
self_reference _avg_sharess	Durchschnittliche Anzahl der Shares aller verlinkten Artikel von Mashable	metrisch	[0,b]
Shares	Anzahl der Shares die der Artikel über alle sozialen Netzwerke erhalten hat	metrisch diskret	[0,a]

Tabelle 19: Variablenbeschreibung, Skalenniveau und Wertebereich für die Variablen der Klasse „Shares“.

Veröffentlichung. Variablen der Kategorie „Veröffentlichung“ geben an, an welchem Wochentag die Artikel veröffentlicht wurden. Da diese Kennzahlen keinen Einfluss auf die Texte als solche haben, werden sie in der Analyse in dieser Arbeit nicht beachtet.

Variablenkürzel	Variablenbeschreibung	Skalenniveau	Wertebereich
Timedelta	Zeitdifferenz in Tagen zwischen Veröffentlichung und Extraktion	metrisch diskret	[0,a]
weekday_is_monday	Der Artikel wurde an einem Montag veröffentlicht	Binär	{0,1}

weekday_is_tuesday	Der Artikel wurde an einem Dienstag veröffentlicht	Binär	{0,1}
weekday_is_wednesday	Der Artikel wurde an einem Mittwoch veröffentlicht	binär	{0,1}
weekday_is_thursday	Der Artikel wurde an einem Donnerstag veröffentlicht	binär	{0,1}
weekday_is_friday	Der Artikel wurde an einem Freitag veröffentlicht	binär	{0,1}
weekday_is_saturday	Der Artikel wurde an einem Samstag veröffentlicht	binär	{0,1}
weekday_is_sunday	Der Artikel wurde an einem Sonntag veröffentlicht	binär	{0,1}
is_weekend	Der Artikel wurde am Wochenende veröffentlicht	binär	{0,1}

Tabelle 20: Variablenbeschreibung, Skalenniveau und Wertebereich für die Variablen der Klasse „Artikelkategorie“.

6.2 Links

http://mashable.com/2013/01/26/infographics-marketing/

http://mashable.com/2013/02/03/super-bowl-ads-live-blog/

http://mashable.com/2013/03/13/ashamalla-sketches-sxsw/

http://mashable.com/2013/03/19/mobile-consumers-infographic/

http://mashable.com/2014/02/13/comcast-time-warner-apple-tv/

http://mashable.com/2014/07/30/jude-law-ad/

http://mashable.com/2014/08/11/content-marketing-roi-data/

http://mashable.com/2014/08/21/microsoft-surface-pro-3-docking-station-review/

http://mashable.com/2014/08/28/qantas-airways-announces-record-loss/

http://mashable.com/2014/09/08/adobe-mobile-benchmark-report/

http://mashable.com/2014/09/08/fran-drescher-shiva-ayyadurai-invented-email/

http://mashable.com/2014/09/15/no-more-meetings/

http://mashable.com/2014/09/16/nyc-camouflage/

http://mashable.com/2014/09/17/skateboarding-in-balloons/

http://mashable.com/2014/09/23/farewell-kings-cross-sydney/

http://mashable.com/2014/09/29/yelp-reviews-national-parks/

http://mashable.com/2014/10/21/ask-dev-frameworks/

http://mashable.com/2014/11/06/gordon-hayward-lebron-james-2/

http://mashable.com/2014/11/06/robot-butler-brandspeak/

http://mashable.com/2014/11/20/gunman-florida-university/

http://mashable.com/2014/11/21/3d-printing-space/

http://mashable.com/2014/12/01/steve-jobs-deposition/

http://mashable.com/2014/12/08/yik-yak-fixes-problem/

7 Quellen

Chris Beier, Daniel Wolfman (2012): http://www.inc.com/chris-beier-and-daniel-wolfman/how-pete-cashmore-founded-mashable.html, abgerufen am 04.02.2016

Chase Peterson-Withorn (2014): http://www.forbes.com/sites/chasewithorn/2014/10/09/how-pete-cashmore-turned-appendicitis-into-a-modern-media-powerhouse/#244ccba1191a, zuletzt abgerufen am 04.02.2016

Chris Raymond (2013): http://www.success.com/mobile/article/pete-cashmore-of-mashable-the-sage-of-social-media, zuletzt abgerufen am 04.02.2016

Fernandes, K., Vinagre, P., & Cortez, P. (2015b): https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity, zuletzt abgerufen am 04.02.2016

De Smedt, T., Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research, 13*, 2031–2035.

Fernandes, K., Vinagre, P., & Cortez, P. (2015a). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In *Progress in Artificial Intelligence* (pp. 535-546). Springer International Publishing.

Williams, Brett, Ted Brown, and Andrys Onsman. "Exploratory factor analysis: A five-step guide for novices." *Australasian Journal of Paramedicine* 8.3 (2012): 1.

Zygmont, C. & Smith, M. R. (2014). “Robust factor an alysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions.” *The Quantitative Methods for Psychology*, 10 (1), 40-55.

Costello, Anna B. & Jason Osborne (2005). “Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis.” *Practical Assessment Research & Evaluation*, 10(7).

Kieffer, Kevin M. "Orthogonal versus Oblique Factor Rotation: A Review of the Literature regarding the Pros and Cons." (1998).

Härdle, Wolfgang, and Léopold Simar. *Applied multivariate statistical analysis*. Vol. 2. Berlin: Springer, 2003.

Backhaus, Klaus, et al. *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Springer-Verlag, 2013.

Peterson, Robert A. "A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis." *Marketing Letters* 11.3 (2000): 261-275.

Galbraith, J. I., Moustaki, I., Bartholomew, D. J., & Steele, F. (2002). *The analysis and interpretation of multivariate data for social scientists*. CRC Press.

Jöreskog, K. G. (2003). Factor analysis by MINRES. *Chicago: Scientific Software. Retrieved June*, *1*, 2005.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, *12*(2), 1-11.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior research methods, instruments, & computers*, *32*(3), 396-402.

Kabacoff, R. I. (2003). Determining the dimensionality of data: a SAS macro for parallel analysis. In *Proceedings of the 28th Annual Meeting of SAS Users Group International. Seattle, WA. Retrieved March* (Vol. 5, No. 2004, pp. 090-28).

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and psychological measurement*, *55*(3), 377-393.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321-327.

Yalcin, I., & Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statistical science*, 275-294.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, *4*(1), 84.

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin*, *101*(1), 126.

Kolenikov, S., & Bollen, K. A. (2012). Testing Negative Error Variances Is a Heywood Case a Symptom of Misspecification? *Sociological Methods & Research*, *41*(1), 124-167.

Pang-Ning Tan, Michael Steinbach, Vipin Kumar: *Introduction to Data Mining.* 1. Auflage, Pearson Verlag, 2005.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999, June). OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod Record* (Vol. 28, No. 2, pp. 49-60). ACM.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Aidong Zhang. Vorlesungsfolien zur Vorlesung CSE 601: Data Mining and Bioinformatics, Herbst 2006. University at Buffalo, The State University of New York. http://www.cse.buffalo.edu/faculty/azhang/cse601/density-based.ppt

Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." *Mining Text Data*. Springer US, 2012. 77-128.

Estivill-Castro, Vladimir. "Why so many clustering algorithms: a position paper." *ACM SIGKDD explorations newsletter* 4.1 (2002): 65-75.

F. Murtag. 1983. Expected time complexity results for hierarchic clustering algorithms which use cluster centres. Information Processing Letters 16 (1983) 237–241

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of marketing research*, 134-148.

Soni, N., & Ganatra, A. (2012). Categorization of Several Clustering Algorithms from Different Perspective: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering* (*IJARCSSE).* Volume 2, Issue 8, 63-68

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer US.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*(15), 3201-3212.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications*, *5*(1), 27-34.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Clustering validity checking methods: part II. *ACM Sigmod Record*, *31*(3), 19-27.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R package for cluster validation. *Journal of Statistical Software (Brock et al., March 2008)*.

Kovács, F., Legány, C., & Babos, A. (2005, November). Cluster validity measurement techniques. In *6th International symposium of hungarian researchers on computational intelligence*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Bock, H. H. (2007). Clustering methods: a history of k-means algorithms. In *Selected contributions in data analysis and classification* (pp. 161-172). Springer Berlin Heidelberg.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15-24.

Hill, T., Lewicki, P., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc..

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010, December). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 911-916). IEEE.

Halkidi, M., & Vazirgiannis, M. (2001). Clustering validity assessment: Finding the optimal partitioning of a data set. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on* (pp. 187-194). IEEE.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer Berlin Heidelberg.

Lei, Xuandong, Xiaoti Hu, and Hongsheng Fang. "Is Your Story Going to Spread Like a Virus? Machine Learning Methods for News Popularity Prediction."

Dumont, F., Macdonald, E. B., & Rincón, A. F. Predicting Mashable and Reddit Popularity Using Closed-form and Gradient Descent Linear Regression.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*, 627-666.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, *53*(4), 675-679.

Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and trends in information retrieval* 2, no. 1-2 (2008): 1-135.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Schnell, R., Hill, P. B., & Esser, E. (2011). *Methoden der empirischen Sozialforschung*. Oldenbourg Verlag.

Przepiórkowski, A., & Ogrodniczuk, M. (Eds.). (2014). *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings* (Vol. 8686). Springer.

Carstensen, K. U., Ebert, C., Ebert, C., Jekat, S., Langer, H., & Klabunde, R. (Eds.). (2009). *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Springer-Verlag.

Goker, A., & Davies, J. (Eds.). (2009). *Information retrieval: searching in the 21st century*. John Wiley & Sons.

Sigbert Klinke, 2015. *Datenanalyse II (Foliensatz multivariate statistics). Folientexte zur Vorlesung.* Wintersemester 2015/2016, Humboldt Universität zu Berlin (Lehrstuhl für Statistik), unveröffentlicht.

Sigbert Klinke, 2014. *Folientexte zur Vorlesung Statistik I.* Sommersemester 2014, Humboldt Universität zu Berlin (Lehrstuhl für Statistik), unveröffentlicht.

Abdi, H. (2003). Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, 792-795.

Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. Pearson Deutschland GmbH.

Lee, C. F., Lee, J. C., & Lee, A. C. (2000). *Statistics for business and financial economics* (Vol. 1, p. 712). Singapore: World Scientific.

Leech, N. L., & Onwuegbuzie, A. J. (2002). A Call for Greater Use of Nonparametric Statistics.

Gibbons, J. D., & Chakraborti, S. (2003). Nonparametric statistical inference fourth edition, revised and expanded. *STATISTICS TEXTBOOKS AND MONOGRAPHS*, *168*.

Klinke, S., Mihoci, A., & Härdle, W. (2010). Exploratory factor analysis in MPlus, R and SPSS. *Invited paper ICOTS8 of the International Association of Statistical Education*.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, *6*(1), 99-106.

Suhr, D. D. (2005). Principal component analysis vs. exploratory factor analysis. *SUGI 30 proceedings*, *203*, 230.

Jolliffe, I. T. (2002). Principal component analysis and factor analysis. *Principal component analysis*, 150-166. (Das ist ein Buchauszug!!!)

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Lorenzo-Seva, U. (2013). How to report the percentage of explained common variance in exploratory factor analysis. *Available ftp: http://psico. fcep. urv*.

Richman, M. B. (1986). Rotation of principal components. *Journal of climatology*, *6*(3), 293-335.

Cleff, T. (2011). *Deskriptive Statistik und moderne Datenanalyse: eine computergestützte Einführung mit Excel, PASW (SPSS) und STATA*. Springer-Verlag.

Mutz, D. C., & Soss, J. (1997). Reading public opinion: The influence of news coverage on perceptions of public sentiment. *Public Opinion Quarterly*, 431-451.

Kiousis, S., Popescu, C., & Mitrook, M. (2007). Understanding influence on corporate reputation: An examination of public relations efforts, media coverage, public opinion, and financial performance from an agenda-building and agenda-setting perspective. *Journal of Public Relations Research*, *19*(2), 147-165.

Iyengar, S., & Kinder, D. R. (2010). *News that matters: Television and American opinion*. University of Chicago Press.

Jung, Y., Park, H., Du, D. Z., & Drake, B. L. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization*, *25*(1), 91-111.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159-179.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

Sugar, C. A., & James, G. M. (2011). Finding the number of clusters in a dataset. *Journal of the American Statistical Association*.

Deepak, P., & Roy, S. (2006). Optics on text data: Experiments and test results.

Hennig, C. (2015). Package 'fpc'.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package 'NbClust'. *J. Stat. Soft*, *61*, 1-36.

Ritter, G. (2014). *Robust cluster analysis and variable selection*. CRC Press.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, *5*(2), 181-204.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Routledge.

Richard L. Gorsuch. *Factor Analysis: Classic Edition.* Psycholoy Press & Routledge Classic Editions.

8 Erklärung zur Urheberschaft

Hiermit erkläre ich, Oliver Brose, dass ich die vorliegende Arbeit allein und unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe. Die Prüfungsordnung ist mir bekannt. Ich habe in meinem Studienfach bisher keine Bachelorarbeit eingereicht.